Multimodal Recognition of Visual Concepts using Histograms of Textual Concepts and Selective Weighted Late Fusion Scheme

Ningning Liu, Emmanuel Dellandréa, Liming Chen, Chao Zhu, Yu Zhang, Charles-Edmond Bichot, Stéphane Bres, Bruno Tellez

▶ **To cite this version:**

Ningning Liu, Emmanuel Dellandréa, Liming Chen, Chao Zhu, Yu Zhang, et al.. Multimodal Recognition of Visual Concepts using Histograms of Textual Concepts and Selective Weighted Late Fusion Scheme. Computer Vision and Image Understanding, 2013, 5, 117, pp.493-512. 10.1016/j.cviu.2012.10.009 . hal-01339139

HAL Id: hal-01339139
https://hal.science/hal-01339139v1

Submitted on 7 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimodal Recognition of Visual Concepts using Histograms of Textual Concepts and Selective Weighted Late Fusion Scheme

Ningning Liu[a,*], Emmanuel Dellandréa[a], Liming Chen[a], Chao Zhu[a], Yu Zhang[a], Charles-Edmond Bichot[a], Stéphane Bres[b], Bruno Tellez[b]

[a] *Université de Lyon, CNRS,*
Ecole Centrale de Lyon, LIRIS, UMR 5205, F-69134, France
[b] *INSA de Lyon, LIRIS, UMR 5205,*
20 Av. A. Einstein, 69621 Villeurbanne Cedex, France

Abstract

The text associated with images provides valuable semantic meanings about image content that can hardly be described by low-level visual features. In this paper, we propose a novel multimodal approach to automatically predict the visual concepts of images through an effective fusion of textual features along with visual ones. In contrast to the classical Bag-of-Words approach which simply relies on term frequencies, we propose a novel textual descriptor, namely the Histogram of Textual Concepts (HTC), which accounts for the relatedness of semantic concepts in accumulating the contributions of words from the image caption toward a dictionary. In addition to the popular SIFT-like features, we also evaluate a set of mid-level visual features, aiming at characterizing the harmony, dynamism and aesthetic quality of visual content, in relationship with affective concepts. Finally, a novel selective weighted late fusion (SWLF) scheme is proposed to automatically select and weight the scores from the best features according to the concept to

*Corresponding author. Tel: (+33)684663045
Email addresses: ningning.liu@ec-lyon.fr (Ningning Liu), emmanuel.dellandrea@ec-lyon.fr (Emmanuel Dellandréa), liming.chen@ec-lyon.fr (Liming Chen), chao.zhu@ec-lyon.fr (Chao Zhu), yu.zhang@ec-lyon.fr (Yu Zhang), Charles-Edmond.Bichot@ec-lyon.fr (Charles-Edmond Bichot), stephane.bres@insa-lyon.fr (Stéphane Bres), bruno.tellez@univ-lyon1.fr (Bruno Tellez)

be classified. This scheme proves particularly useful for the image annotation task with a multi-label scenario. Extensive experiments were carried out on the MIR FLICKR image collection within the ImageCLEF 2011 photo annotation challenge. Our best model, which is a late fusion of textual and visual features, achieved a MiAP (Mean interpolated Average Precision) of 43.69% and ranked 2^{nd} out of 79 runs. We also provide comprehensive analysis of the experimental results and give some insights for future improvements.

Keywords:
Image classification, textual feature, visual feature, fusion, ImageCLEF photo annotation.

---

1. Introduction

Machine-based recognition of visual concepts aims at recognizing automatically from images high-level semantic concepts (HLSC), including scenes (indoor, outdoor, landscape, *etc.*), objects (car, animal, person, *etc.*), events (travel, work, *etc.*), or even emotions (melancholic, happy, *etc.*). It proves to be extremely challenging because of large intra-class variations (clutter, occlusion, pose changes, *etc.*) and inter-class similarities [1, 2, 3, 4]. The past decade has witnessed tremendous efforts from the research communities as testified the multiple challenges in the field, *e.g.*, Pascal VOC [5], TRECVID [6] and ImageCLEF [7, 8, 9, 10]. Most approaches to visual concept recognition (VCR) have so far focused on appropriate visual content description, and have featured a dominant Bag-of-Visual-Words (BoVW) representation along with local SIFT descriptors. Meanwhile, increasing works in the literature have discovered the wealth of semantic meanings conveyed by the abundant textual captions associated with images [11, 12, 13]. As a result, multimodal approaches have been increasingly proposed for VCR by making joint use of user textual tags and visual descriptions to bridge the gap between low-level visual features and HLSC. The work presented in this paper is in that line and targets an effective multimodal approach for VCR.

1.1. Related works

The state of the art for VCR using only visual content has proposed a large set of local visual features, including SIFT [14], Color SIFT [15], HOG [16], DAISY [17], LBP [18], Color LBP [19], which are all based on the first order gradient information. The dominant approach for visual content

representation is the BoVW method [20], which represents an image as an orderless collection of local visual features extracted from a dense grid or sparse keypoints over the image. An image can thus be described by a histogram, using a hard or soft assignment over a visual dictionary of fixed size learnt from a training dataset. While this approach has largely demonstrated its effectiveness in various challenges for VCR such as Pascal VOC [5] and is also the prevalent technique used in the ImageCLEF Image Annotation task [9], its major shortcoming is still its lack of descriptive power as regard to HLSCs because of its nature of low-level features.

There is an increasing interest in capturing emotional and aesthetic aspects of visual content, including features based on the experimentally determined color factors [21, 22], texture attributes [23], shape elements [24, 25] as well as aesthetic features [26, 27]. Meanwhile, all the emotion related features so far proposed were only evaluated on rather small and specifically tailored datasets. In this work, we evaluated the usefulness of these features along with the aesthetic ones for the recognition of affective concepts as well as other general visual concepts on a large and general image dataset.

The BoVW approach actually originates from the field of information retrieval where a text document is often represented as a Bag-of-Words (BoW) and described according to the vector space model [28] as a vector of terms, each component of which is a kind of word count or term frequency as exemplified by TF-IDF (Term Frequency-Inversed Document Frequency). This model has undergone several extensions, including latent semantic analysis (LSA) [29], probabilistic LSA [30] and Latent Dirichlet allocation (LDA) [31]. The major drawback of these word frequency statistic-based approaches is their lack of semantic sensitivity, for two reasons. First, a text document is simply interpreted as an unordered collection of words, thus disregarding grammar and even word order; second, a text document is further summarized as a vector of term frequencies, thereby failing to capture the relatedness between words. The literature has recorded a number of attempts trying to remedy these two shortcomings, including in particular the use of linguistic structures [32], *e.g.*, compound terms such as *operating system*, binary relations such as subject-verb or verb-object, *etc.*, or distributional term representations (DTRs) [33] which propose to characterize the meaning of a term from its context, *i.e.* the other terms with which it frequently co-occurs within a window or simply the documents in which it frequently occurs. While various essays using the former approach prove to be surprisingly ineffective both for the tasks of information retrieval and text categorization

as compared to the BoW approach [34], the latter clearly represents a step forward towards taking into account the relatedness of terms through their context [35]. But still, the component of a context vector is simply a term frequency count either within a document or within a window of a given size. It is interesting to note a very recent move which tries, through these DTRs, to capture the relatedness of terms occurring in the textual bimodalities of image captions and labels for the purpose of image retrieval [36].

As current BoVW based works seem to be reaching the performance ceiling for VCR, there exists an increasing interest in multimodal approaches [11, 37, 38], attempting to make joint use of visual descriptions and abundant tags associated with images for better prediction of visual concepts. Still, the dominant approach for characterizing the textual content of image tags is the vector space model, using different variants of term counts or frequencies after some basic preprocessing, *e.g.*, stop words removal, stemming, *etc.* All these works consistently demonstrate that the textual features can improve the performance of VCR when used jointly with visual features. Meanwhile, as these textual features are mostly term counts-based, they fail to capture the relatedness between semantic concepts.

As far as multimodal approaches are concerned, they require a fusion strategy to combine information from multiple sources, *e.g.*, visual stream and sound stream for video analysis [39, 40], textual and visual content for multimedia information retrieval [41, 42], *etc.* This fusion can be carried out at feature level (called *early fusion*) [43] or at score level (called *late fusion*) [44], or even at some intermediate levels, *e.g.*,kernel level [38, 37]. While early fusion is straightforward and simply consists of concatenating the features extracted from various information sources into a single representation, its disadvantage is also well known: the curse of dimensionality and the difficulty in combining features of different natures into a common homogeneous representation. As a result, late fusion strategies, which consist of integrating the scores as delivered by the classifiers on various features through a fixed combination rule, *e.g.*, sum, are competitive alternatives in the literature [45, 46]. They not only provide a trade-off between preservation of information and computational efficiency but also prove to perform favorably as compared to early fusion methods in several comparative studies, *e.g.*, [47, 39] on visual concept detection in video sequences. Furthermore, a comprehensive comparative study of various combination rules, *e.g.*, sum, product, max, min, median, and majority voting, by Kittler *et al.* [48], suggests that the sum rule is much less sensitive to the error of individual

classifiers when estimating posterior class probability. The proposed fusion scheme, Selective Weighted Late Fusion (SWLF), falls into the category of late fusion strategies and selectively chooses the best classifiers to optimize the overall mean average precision.

1.2. The proposed approach

In this paper, we propose a novel multimodal approach for VCR that builds on a novel textual representation along with visual features through an effective selective weighted late fusion scheme (SWLF). In contrast to term frequency-based text representations mostly used in VCR, the proposed novel textual representation, namely the Histogram of Textual Concepts (HTC), captures the relatedness of semantic concepts while SWLF automatically selects and weights the best discriminative features for each visual concept to be predicted in optimizing the overall mean average precision. Furthermore, we also propose to study the usefulness of a set of mid-level features, *e.g.*, emotion and aesthetics related ones, for the recognition of sentiment concepts as well as other general visual ones. The proposed approach was extensively evaluated on the MIR FLICKR image collections [7, 8] within the ImageCLEF 2011 photo annotation challenge and demonstrates its effectiveness. Indeed, our best prediction model, which is a late fusion of the novel HTCs and visual features through SWLF, achieved a MiAP (Mean interpolated Average Precision) of 43.69% and ranked the 2^{nd} best performance out of 79 runs.

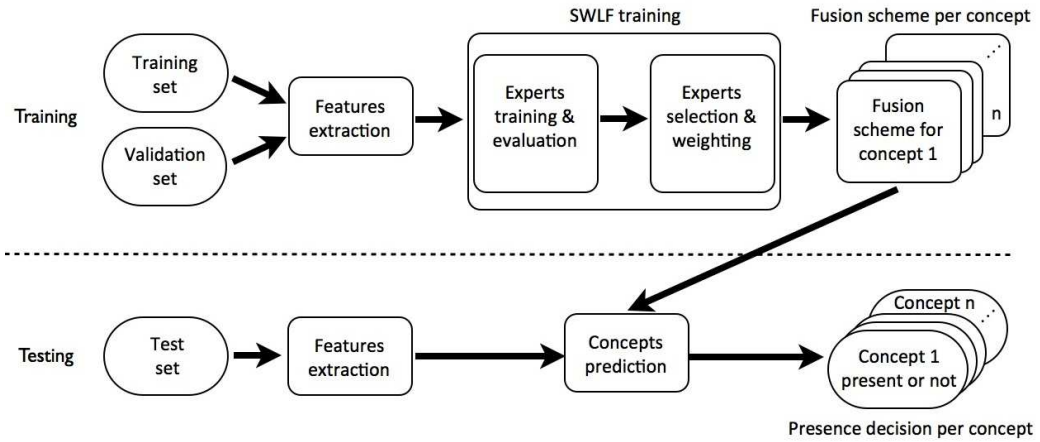


Figure 1: The flowchart of the proposed approach for visual concept recognition.

Figure 1 depicts the flowchart of the proposed approach for VCR which mainly includes two stages: a training stage and a testing stage. The training stage consists of training experts through SVM for each pair of concept and type of features using a training set. These experts are then evaluated using a validation set to learn SWLF. The testing stage proceeds to extract various types of features from an input image, and then to apply the corresponding fusion scheme learnt by SWLF for each concept to deliver a recognition decision. The contributions of this work are threefold and can be summarized as follows:

- A novel effective textual feature, HTC, is proposed to capture the relatedness of semantic concepts and accounts for the sparsity of image tags. Several variants of HTC are also provided and compared, using two different semantic word similarities and two different dictionaries, including in particular an English word-based affective database ANEW [49].

- We investigate a set of mid-level features, which are related to harmony, dynamism, aesthetic quality, emotional color representation, *etc.*, and evaluate their efficiency for the specific problem of affective concepts classification.

- We propose a novel SWLF scheme which selects the best features and weights their scores for each concept. This fusion scheme proves particularly efficient for fusing visual and textual modalities in comparison with some other standard fusion schemes including min, max, and mean.

The rest of this paper is organized as follows. The novel textual feature HTC is introduced in Section 2. Section 3 presents the set of mid-level visual features along with some popular low-level features such as color, texture, shape and local descriptors. The fusion strategy is investigated in Section 4. The experimental results are analyzed in Section 5. Finally, section 6 draws the conclusion and gives some hints for future work.

2. Textual features

The last few years have seen an impressive growth of sharing websites particularly dedicated to videos and images. The famous Flickr website[1] for example, from which is extracted the MIR FLICKR image collection that we investigate in this paper, allows users to upload and share their images and to provide a textual description under the form of tags or legends. These textual descriptions are a rich source of semantic information on visual data that is interesting to consider for the purpose of VCR or multimedia information retrieval. However,while there exist abundant captioned images on the Internet, a textual caption for a given image is generally very sparse (8.7 tags on average per image in MIR FLICKR). An example is given in Figure 2 where a picture of a peacock is associated with user tags such as "bird", "beautiful", "interestingness". In this section, we first introduce a novel descriptor of textual content, namely Histograms of Textual Concepts, then present ten variants using different dictionaries, semantic similarity measurements and accumulating operators.

{0A432C9F-1732-45E6-90F7-A6A7B75FA889}.jpg

Flickr user tags: peacock, bird, beautiful, pretty, feathers, waimea, waimeafalls, explore, animal, interestingness

Figure 2: An example image with sparse Flickr user tags, including however semantic concepts, *e.g.*, "bird", "beautiful", "interestingness", *etc.*

[1] http://www.flickr.com/

2.1. HTC: a Histogram of Textual Concepts

We have seen that the dominant BoW approach fails to describe the fineness and the relatedness of semantic concepts. Indeed, the BoW kind approaches assume that word terms are basically statistically independent, thereby mismatching text documents close in content but with different term vocabulary. In contrast, we propose the Histograms of Textual Concepts (HTC) to capture the semantic relatedness of concepts. HTC is inspired from a model that we can call *componential space model*, such as conceptual vector [50], which describes the meaning of a word by its atoms, components, attributes, behavior, related ideas, *etc.* For instance, the concept of "rain" can be described by "water", "liquid", "precipitation", "dripping liquid", "monsoon", *etc.* thus in a much similar way when users tag photos. Similarly, the concept "peacock" as illustrated in Figure 2 can be described by "bird", "male", "beautiful", "pretty", "feathers", "plumage", "animal", *etc.*

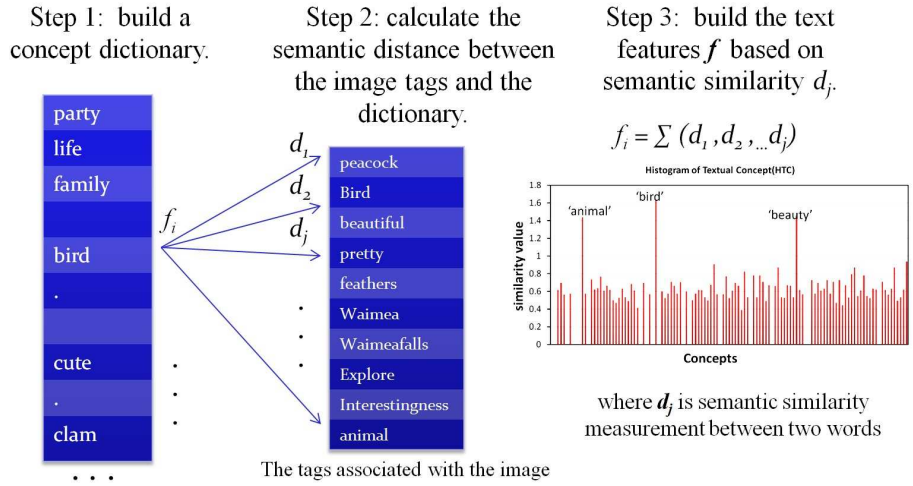


Figure 3: The three steps process of our HTC algorithm (taking Figure 2 as an input example.)

Specifically, the HTC of a text document is defined as a histogram of textual concepts towards a vocabulary or dictionary, and each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of the contribution of each word within the text document toward the underlying concept according to a predefined semantic similarity measure. Given a dictionary D and a semantic similarity measurement S, HTC can be simply extracted from the tags of an image through a three-step

process as illustrated in Figure 3. Note that the tags such as "peacock", "bird", "feathers", "animal" all contribute to the bin values associated with the "animal" and "bird" concepts according to a semantic similarity measurement whereas the tags such "beautiful", "pretty", "interestingness" all help peak the bin value associated with the concept "cute". This is in clear contrast to the BoW approaches where the relatedness of textual concepts is simply ignored as word terms are statistically counted. The algorithm for the extraction of a HTC feature is detailed in Algorithm 1.

Algorithm 1: Histogram of Textual Concepts (HTC)

Input: Tag data $W = \{w_t\}$ with $t \in [1, \mathtt{T}]$, dictionary $D = \{d_i\}$ with $i \in [1, \mathtt{d}]$.
Output: Histogram f composed of values f_i with $0 \leq f_i \leq 1$, $i \in [1, \mathtt{d}]$.

- Preprocess the tags by using a stop-words filter.
- If the input image has no tags ($W = \emptyset$), return f with $\forall i \ f_i = 0.5$.[1]
- Do for each word $w_t \in W$:
  1. Calculate $dist(w_t, d_i)$, where $dist$ is a semantic similarity distance between w_t and d_i.
  2. Obtain the semantic matrix S as: $S(t, i) = dist(w_t, d_i)$.
- Calculate the feature f as: $f_i = \sum_{t=1}^{\mathtt{T}} S(t, i)$, and normalize it to [0 1] as: $f_i = f_i / \sum_{j=1}^{\mathtt{d}} f_j$.

[1] When an input image has no tag at all, in this work we simply assume that every bin value is 0.5, therefore at halfway between a semantic similarity measurement 0 (no relationship at all with the corresponding concept in the dictionary) and 1 (full similarity with the corresponding concept in the dictionary). Alternatively, we can also set these values to the mean of HTCs over the captioned images of a training set.

The advantages of HTC are multiple. First, for a sparse text document as image tags, HTC offers a smooth description of the semantic relatedness of user tags over a set of textual concepts defined within the dictionary. More importantly, in the case of polysemy, HTC helps disambiguate textual concepts according to the context. For instance, the concept of "bank" can refer to a financial intermediary but also to the shoreline of a river. However, when a tag "bank" comes with a photo showing a financial institution, correlated tags such as "finance", "building", "money", *etc.*, are very likely to be used, thereby clearly distinguishing the concept "bank" in finance from that of a river where correlated tags can be "water", "boat", "river", *etc.* Similarly,

in the case of synonyms, the HTC will reinforce the concept related to the synonym as far as the semantic similarity measurement takes into account the phenomenon of synonyms.

2.2. Variants of HTC

The computation of HTC requires the definition of a dictionary and a proper semantic relatedness measurement over textual concepts. In this work, we compare the use of two dictionaries. The first one, namely *D_99*, is a dictionary composed of the 99 visual concepts to be detected within the photo annotation task of ImageCLEF 2011, while the second one, *D_Anew*, is the set of 1034 English words used in the ANEW study [49]. The interest of the ANEW dictionary lies in the fact that each of its word is rated on a scale from 1 to 9 using affective norms in terms of *valence* (affective dimension expressing positive versus negative), *arousal* (affective dimension expressing active versus inactive) and *dominance* (affective dimension expressing dominated versus in control). For instance, according to ANEW, the concept "beauty" has a mean valence of 7.82, a mean arousal of 4.95 and a mean dominance of 5.23 while the concept "bird" would have a mean valence of 7.27, a mean arousal of 3.17 and a mean dominance of 4.42. Therefore, *D_99* allows for the projection of all user tags into the space of 99 visual concepts to be classified whereas *D_Anew* seems better armed for the 9 sentiment concepts, *e.g.*, melancholic, happy, active, *etc.* which were newly introduced in the photo annotation task within ImageCLEF 2011.

Using the affective ratings of the ANEW concepts and the HTCs computed over image tags, one can further define the coordinates of an image caption in the three dimensional affective space [51], in terms of valence, arousal and dominance by taking a linear combination of the ANEW concepts weighted by the corresponding HTC values. More precisely, given a HTC descriptor f extracted from a text document, the valence, arousal and dominance coordinates of the text document can be computed as follows:

$$f_{valence} = (1/\mathtt{d}) \sum_i (f_i * V_i) \tag{1}$$

$$f_{arousal} = (1/\mathtt{d}) \sum_i (f_i * A_i) \tag{2}$$

$$f_{dominance} = (1/\mathtt{d}) \sum_i (f_i * D_i) \tag{3}$$

where V_i, A_i and D_i are respectively the valence, the arousal and the dominance of the i^{th} word w_i in the *D_Anew* dictionary, and $\mathtt{d}$ is the size of *D_Anew*. However, Bradley *et al.* [52] pointed out that the measurement of dominance information is not stable. As a result, we only made use of Equations (1) and (2) defined as a variant of HTC features in our participation to the ImageCLEF 2011 photo annotation task.

We also implement and compare two measurements of semantic similarities between two textual concepts, namely the *path* and the *wup* distances [53] that are based on the WordNet ontology [54]. Given two synsets w_1 and w_2, the *path* and the *wup* distances are defined by:

$$d_{path}(w_1, w_2) = \frac{1}{1 + spl(w_1, w_2)} \tag{4}$$

$$d_{wup}(w_1, w_2) = \frac{2 \times depth(lcs(w_1, w_2))}{depth(w_1) + depth(w_1)} \tag{5}$$

where $lcs(w_1, w_2)$ denotes the least common subsumer (most specific ancestor node) of the two synsets w_1 and w_2 in the WordNet taxonomy, $depth(w)$ is the length of the path from w to the taxonomy root, and $spl(w_1, w_2)$ returns the distance of the shortest path linking the two synsets (if one exists). Note that the *path* and the *wup* measurements have opposite polarity. When the two synsets w_1 and w_2 are identical, *path* returns 1 while *wup* returns 0. Therefore, when using *wup* for accumulating the semantic similarities in the computation of HTC, its polarity is first changed to a positive one in our work.

Finally, for comparison purpose, in addition to the *sum* operator which accumulates the semantic relatedness of the tags of an image toward a predefined dictionary, we also make use of the *max* operator which handles the semantic similarities by keeping only the maximal value of all image tags toward each concept in the dictionary. In this case, the accumulation of the semantic relatedness $f_i = \sum_t S(t, i)$ in the HTC computation is replaced by $f_i = \max_t S(t, i)$.

These different variants of HTC are listed in Table 1. In this table, the feature names are related to the way they are computed according to the aforementioned alternatives. For instance, *txtf_99ps* refers to the HTC variant using the dictionary *D_99* made of ImageCLEF 2011 concepts along with

Table 1: Different variants of the textual features based on HTC.

Feature name	Dictionary	Similarity measure	Accumulating method
txtf_99ps	*D_99*	path	$f_i = \sum_t S(t,i)$
txtf_99pm	*D_99*	path	$f_i = \max_t S(t,i)$
txtf_99ws	*D_99*	wup	$f_i = \sum_t S(t,i)$
txtf_99wm	*D_99*	wup	$f_i = \max_t S(t,i)$
txtf_1034ps	*D_Anew*	path	$f_i = \sum_t S(t,i)$
txtf_1034pm	*D_Anew*	path	$f_i = \max_t S(t,i)$
txtf_1034ws	*D_Anew*	wup	$f_i = \sum_t S(t,i)$
txtf_1034wm	*D_Anew*	wup	$f_i = \max_t S(t,i)$
txtf_1034pva	*D_Anew*	path	$f_i = \max_t S(t,i)$
txtf_1034wva	*D_Anew*	wup	$f_i = \max_t S(t,i)$
txtf_1034pvad	*D_Anew*	path	$f_i = \max_t S(t,i)$
txtf_1034wvad	*D_Anew*	wup	$f_i = \max_t S(t,i)$

the *path* distance as semantic similarity measurement, and the *sum* accumulating operator. *txtf_1034pvad* refers to the valence, arousal and dominance coordinates, namely $f_{valence}$, $f_{arousal}$ and $f_{dominance}$, which are computed using Equations (1), (2) and (3) while the underlying HTC variant is computed using ANEW vocabulary *D_Anew* and the *path* distance.

3. Visual features

To describe the visual content of an image, we also follow the dominant BoVW approach which views an image as an unordered distribution of local image features extracted from salient image points, called "interest points" [14, 55] or more simply from points extracted on a dense grid [56, 57]. In this work, we make use of several popular local descriptors, including C-SIFT, RGB-SIFT, HSV-SIFT [15] and DAISY [58], extracted from a dense grid [17]. An image is then modelled as a BoVW using a dictionary of 4000 visual words and hard assignment. The codebook size, 4000 in this work, results from a tradeoff between computational efficiency and the performance over a training dataset. The visual words represent the centers of the clusters obtained from the k-means algorithm.

Meanwhile, in order to capture the global ambiance and layout of an image, we further compute a set of global features, including descriptions of color information in the HSV color space in terms of means, color histograms and color moments, textures in terms of LBP [18], Color LBP [19], co-occurrence and auto-correlation, as well as shape information in terms of histograms of line orientations quantized into 12 different orientations and computed by the Hough transform [59].

In addition to these local and global low-level features, we also collect and implement a set of mid-level features [60, 25, 24] which are mostly inspired from studies in human visual perception, psychology [21], cognitive science, art [61], *etc.*, thus in close relationships with the 9 sentiment concepts newly introduced in the image annotation task at ImageCLEF 2011. These mid-level features include emotion related visual features, aesthetic and face related features.

Emotion related features.

Color. Valdez and Mehrabian [21] carried out psychological experiments and evidenced significant relationships between color saturation and brightness, and emotion dimensions. They further expressed these relationships in terms of pleasure (valence), arousal and dominance axis according to the following equations:

$$Pleasure = 0.69V + 0.22S \tag{6}$$

$$Arousal = -0.31V + 0.60S \tag{7}$$

$$Dominance = 0.76V + 0.32S \tag{8}$$

where S and V refer to the mean value of brightness and saturation in HSV color space, respectively.

Texture. Tamura *et al.* [23] proposed a set of texture features strongly correlated with human visual perception [23] and proved successful for affective image classification [62, 24]. Therefore, in this work we implemented Tamura features including coarseness, contrast, directionality.

Harmony. Itten [61] has shown that color combinations can produce effects such as harmony, non-harmony. Indeed, visual harmony can be obtained by combining hues and saturations so that an effect of stability on the human eye can be produced. The presence of harmonious colors convey stability and joy whereas the presence of non complementary colors gives an

anxious and nervous feeling. This harmony can be represented using the Itten sphere where contrasting colors have opposite coordinates [22, 63, 25, 24]. In the case of harmonious colors, color positions on the sphere are connected thanks to regular polygons. The corresponding harmony feature is thus built by identifying dominant colors and plotting them into the color sphere. Then, the polygon linking these colors is characterized by a value in such a way that a value close to 1 corresponds to a regular polygon whose center is next to the sphere center, thereby characterizing a harmonious image, while a value close to 0 corresponds to an irregular polygon characterizing a non harmonious image.

Dynamism. Lines within an image also convey important meanings [63, 25, 24]. Indeed, oblique lines communicate dynamism and action whereas horizontal or vertical lines rather communicate calmness and relaxation. This can be combined with colors in order to produce complex effects suggesting particular feelings or emotions to the viewer. Therefore, to characterize dynamism in images, the ratio is computed between the numbers of oblique lines (detected by a Hough Transform [64]) with respect to the total number of lines in an image.

Aesthetic qualities. Aesthetics in photographs refers to the feeling of the beauty perceived by people. An image of good aesthetic quality usually induces a pleasant experience. Therefore, we have implemented several aesthetic features proposed by R. Datta *et al.* [26] and some others from Y.Ke *et al.* [27]. We expect them to be useful for identifying some affective concepts, for their relatedness to arts and feelings.

Face related features. Finally, we also implemented a face counting method according to Viola and Jones face detector [65]. The rational is that the knowledge of the number of faces within an image can give useful clues to characterize some concepts that involve the human presence such as "person", "portrait", "family", *etc.*

Table 2 summarizes all the visual features that we have implemented for the purpose of VCR.

Table 2: Summary of the visual features.

Category	Short name	#	Short Description
Color	grey_hist	128	128-bin histogram computed from the grey level image.
	color_hsv	132	Concatenation of the 64-bin histograms computed from each HSV channel.
	color_ moment	144	3 central moments (mean, standard deviation and skewness) on HSV channels using a pyramidal image representation.
	color_mSB	2	Mean saturation and brightness in HSV color space.
Texture	texture_ lbp	256	Standard LBP features [18].
	texture_ tamura	3	Tamura features [23] including coarseness, contrast, directionality.
	texture_ cooccu	16	Distribution of co-occurring values in the image at a given offset [66].
	texture_ autocorr	132	Autocorrelation image coefficients [67].
	hsvLbp invLbp rgbLbp oppoLbp	1311 (each)	Four multi-scale color LBP operators based on different color spaces [19].
Shape	shape_ histLine	12	Histogram of 12 different orientations by using Hough transform [59].
Local descriptors	c-sift rgb-sift hsv-sift oppo-sift	4000 (each)	Four SIFT descriptors based on different color spaces and computed on a dense grid [15, 68].
	daisy	4000	DAISY descriptor computed on a dense grid [17].
Mid-level	mlevel_ PAD	3	Emotional coordinates based on HSV color space according to [21].
	mlevel_ harmony	1	Color harmony of images based on Itten's color theory [61].
	mlevel_ dynamism	1	Ratio between the numbers of oblique lines in images with respect to the total number of lines [22, 63].
	mlevel_ aesthetic YKe	5	Y.Ke *et al.* [27] aesthetic criteria including: spatial distribution of edges, hue count, blur, contrast and brightness .
	mlevel_ aesthetic Datta	44	Most of the features (44 of 56) except those that are related to IRM (integrated region matching) technique [26].
	mlevel_ facect	5	Number of faces in the image detected by using 5 different pose configurations of the face detector from [65].

4. Selective Weighted Late Fusion

The fusion scheme that we implement is a selective weighted late fusion (SWLF), which shares the same idea as the adaptive score level fusion scheme proposed by Soltana *et al.* [69]. While a late fusion at score level is reputed as a simple and effective way to fuse features of different nature for machine-learning problems, the proposed SWLF builds on two simple insights. First, the score delivered by a feature type should be weighted by its intrinsic quality for the classification problem at hand. Second, in a multi-label scenario where several visual concepts may be assigned to an image, different visual concepts may require different features which best recognize them. For instance, the "sky" concept may greatly require global color descriptors, while the best feature to recognize a concept like street could be a segment-based feature for capturing straight lines of buildings. The whole SWLF framework is illustrated in Figure 4.

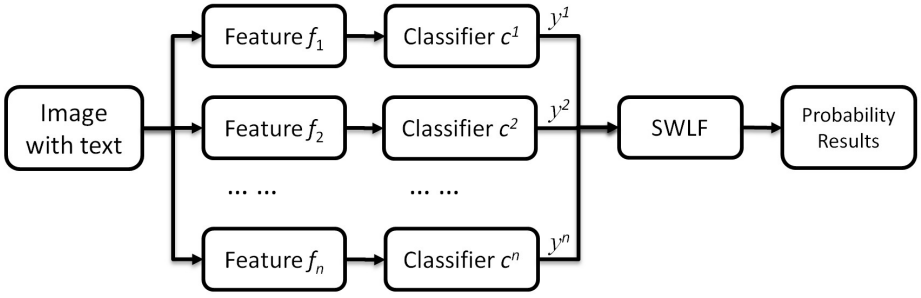


Figure 4: The framework of the SWLF scheme. For each image and each concept, the associated tags are analysed to extract the textual features for textual classifiers. Meanwhile, visual features are extracted to feed visual classifiers. Experts (classifiers) are then combined to predict the presence of a given concept in the input image.

Specifically, the SWLF scheme is implemented as follows. The training dataset is first divided into two parts composed of a training set and a validation set. For each visual concept, a binary classifier (one versus all) is trained, which is also called *expert* in the subsequent, for each type of features using the data in the training set. Thus, for each concept, we generate as many experts as the number of different types of features. The quality of each expert can then be evaluated through a quality metric using the data in the validation set. In this work, the quality metric is chosen to be the interpolated Average Precision (iAP). In this case, the higher the iAP is for

a given expert, the more weight should be given to the score delivered by that expert in the sum of weighted scores for a late fusion. Concretely, given a visual concept k, the quality metrics, *i.e.* iAP, produced by all the experts are first normalized into w_k^i. To perform a late fusion of all these experts at score level, the *sum of weighted scores* is then computed according to Equation (9):

$$score : z_k = \sum_{i=1}^{N} (w_k^i * y_k^i), \tag{9}$$

where y_k^i represents the score of the i^{th} expert for the concept k, and w_k^i stands for the normalized iAP performance of the feature f_i on the validation dataset. In the subsequent, late fusion through Equation (9) is called *weighted score* rule.

For the purpose of comparison, we also consider three other score level fusion schemes, namely "min", "max" or "sum" rules that are recalled respectively in equations (10), (11), (12):

$$min : z_k = min(y_k^1, y_k^2, ..., y_k^N); \tag{10}$$

$$max : z_k = max(y_k^1, y_k^2, ..., y_k^N); \tag{11}$$

$$mean : z_k = \frac{1}{N} \sum_{i=1}^{N} y_k^i; \tag{12}$$

Actually, these three fusion rules can have very simple interpretation. The *min* fusion rule is the consensus voting. A visual concept is recognized only if all the experts recognize it. The *max* rule can be called alternative voting. A visual concept is recognized as long as one expert has recognized it. Finally, the *mean* rule can be assimilated as the majority voting where a concept is recognized if the majority of the experts recognize it.

In practice, one discovers that the late fusion of all the experts leads to a decrease in the global classification accuracy, *i.e.* the mean iAP over the whole set of visual concepts to be recognized. The reason could be that some of the features so far proposed can be noisy and irrelevant to a certain number of visual concepts, thus disturbing the learning process and lowering the generalization skill of the learnt expert on the unseen data. For this purpose,

we further implement the SWLF scheme based on a wrapper feature selection method, namely the SFS method (Sequential Forward Selection) [70], which firstly initializes an empty set, and at each step the feature that gives the highest correct classification rate along with the features already included is added to the set of selected experts to be fused. More specifically, for each visual concept, all the experts are sorted in a decreasing order according to their iAP. At a given iteration N, only the first N experts are used for late fusion and their performances are evaluated over the data of the validation set. N keeps increasing until the overall classification accuracy measured in terms of MiAP starts to decrease. The procedure of the SWLF algorithm is detailed in Algorithm 2.

Several variants of SWLF are conceivable. For example, instead of fixing the same number of experts N for all concepts, it is possible to select the number of experts on a per-concept basis. Thus the number of experts can be different for each concept. Another variant concerns the way the experts are selected at each iteration. Indeed, instead of adding the n^{th} best expert at iteration n to the set of previously selected $n-1$ experts, one can also select the expert which yields the best combination of n experts, in terms of $MiAP$, once added to the set of $n-1$ experts already selected at the previous iteration.

As a late fusion strategy, the computational complexity of SWLF can be computed in terms of the number of visual concepts, K and the number of types of features, M. This complexity is $O(K \times M^2)$. Note that the optimized fusion strategy achieved through SWLF only needs to be trained once on the training and validation datasets.

SWLF combines an ensemble of experts for a better prediction of class labels, *i.e.* visual concepts in this work. From this regard, SWLF can also be viewed as a method of ensemble learning [71] which aims to use multiple models to achieve better predictive performance than could be obtained from any of the constituent models. Nevertheless, SWLF differs from popular bagging methods [72], *e.g.* random forest, which involve having each expert in the ensemble trained using a randomly drawn subset of a training set and vote with equal weight. In the case of SWLF, the training dataset is divided into a training set and a validation set which are used to train experts and SWLF to select the best ones for fusing using different weights.

Algorithm 2: Selective Weighted Late Fusion (SWLF)

Input: Training dataset T (of size N_T) and validation dataset V (of size N_V).
Output: Set of N experts for the K concepts $\{C_k^n\}$ and the corresponding set of weights $\{\omega_k^n\}$ with $n \in [1, N]$ and $k \in [1, K]$.
Initialization: $N = 1$, $MiAP_{max} = 0$.

- Extract M types of features from T and V
- For each concept $k = 1$ to K
  - For each type of feature $i = 1$ to M
    1. Train the expert C_k^i using T
    2. Compute ω_k^i as the iAP of C_k^i using V
  - Sort the ω_k^i in descending order and denote the order as j^1, j^2, ..., j^M to form $W_k = \{\omega_k^{j^1}, \omega_k^{j^2}, ..., \omega_k^{j^M}\}$ and the corresponding set of experts $E_k = \{C_k^{j^1}, C_k^{j^2}, ..., C_k^{j^M}\}$
- For the number of experts $n = 2$ to M
  - For each concept $k = 1$ to K
    1. Select the first n experts from E_k : $E_k^n = \{C_k^1, C_k^2, ..., C_k^n\}$
    2. Select the first n weights from W_k : $W_k^n = \{\omega_k^1, \omega_k^2, ..., \omega_k^n\}$
    3. For $j = 1$ to n : Normalise $\omega_k^{j'} = \omega_k^j / \sum_{i=1}^n \omega_k^i$
    4. Combine the first n experts into a fused expert, using the *weighted score* rule through Equation (9): $z_k = \sum_{j=1}^n \omega_k^{j'} . y_k^j$ where y_k^j is the output of C_k^j
    5. Compute $MiAP_k^n$ of the fused expert on the validation set V
  - Compute $MiAP = 1/K . \sum_{k=1}^K MiAP_k^n$
  - If $MiAP > MiAP_{max}$
    * Then $MiAP_{max} = MiAP$, $N = n$
    * Else break

5. Experimental evaluation

We carried out extensive experiments on the MIR FLICKR image collection [7, 8] that was used within the ImageCLEF 2011 photo annotation challenge [9]. The database is a subset of MIR FLICKR-1M image collection from thousands of the real world users under a creative common license. The participants of the challenge were asked to elaborate methods in order to automatically annotate a test set of 10,000 images with 99 visual concepts (including 9 new emotional concepts) [73]. A training set of 8,000 images was

provided. The task could be solved using three different types of approaches [9]:

- Visual: automatic annotation using visual information only.
- Textual: automatic annotation using textual information only (Flickr user tags and image metadata).
- Multimodal: automatic multimodal annotation using visual information and/or Flickr user tags and/or EXIF information.

The performance was quantitatively measured by the Mean interpolated Average Precision (MiAP) as the standard evaluation measure, while the example-based evaluation applies the example-based F-Measure (F-Ex) and Semantic R-Precision (SRPrecision) [9]. In this paper, we focus on the evaluation using MiAP.

In this section, we investigate the proposed approach under the following conditions: (1) the performance of the visual modality using only visual features; (2) the performance of the textual modality using only textual features; (3) the effect of combining textual and visual features through our SWLF scheme; (4) the usefulness of the set of affect related features for the recognition of the 9 emotional concepts; (5) the performance of our approaches in the photo annotation task at ImageCLEF 2011; (6) discussion of the usefulness of the proposed textual HTC features to the overall performance of our participation to ImageCLEF 2011 photo annotation task and the generalization skill of the fused experts. We start by describing the experimental setup.

5.1. Experimental setup

The initial training dataset, provided by ImageCLEF 2011 for the photo annotation task, was first divided into a training set (50%, 4005 images) and a validation set (50%, 3995 images), and balanced the positive samples of most concepts as half for training and half for validation. These subsets remain the same for all the following experiments. The proposed features, both textual and visual, were then extracted from the training and validation sets. The Support Vector Machines (SVM) [74] were chosen as classifiers (or experts) for their effectiveness both in terms of computation complexity and classification accuracy. A SVM expert was trained for each concept and each type of features, as described in Section 4. Following J. Zhang *et al.* [75], we used χ^2 kernel for visual histogram-based features (including color histogram, color_hsv histogram, LBP-based, SIFT-based, DAISY) and RBF kernels for the other features. The RBF and χ^2 kernel functions are defined by:

$$K_{rbf}(F, F') = exp^{-\frac{1}{2\sigma^2}\|(F-F')\|^2} \tag{13}$$

$$K_{\chi^2}(F, F') = exp^{\frac{1}{I}\sum_{1=1}^{n}\frac{(F_i-F'_i)^2}{F_i+F'_i}} \tag{14}$$

where F and F' are the feature vectors, n is their size, I is the parameter for normalizing the distances which was set at the average value of the training set, and σ was set at $\sqrt{n/2}$.

We made use of the LibSVM library [76] as the SVM implementation (C-Support Vector Classification). The tuning of different parameters for each SVM expert was performed empirically according to our experiments, in which the weight of negative class ("-w-1") was set at 1, and the weight for positive class ("-w1") was optimized on the validation set using a range of 1 through 30.

The test dataset, whose labels were not available at the time of our submission to ImageCLEF 2011 photo annotation task, is composed of 10000 images.

5.2. Experimental results on the visual modality

The visual features were studied in Section 3. There are 24 different types of visual features as synthesized in Table 2. Figure 5 shows the performance of each type of visual features on the validation set. As we can see, SIFT like local features (RGB-SIFT, OPPO-SIFT, C-SIFT, HSV-SIFT) are the most effective ones among all the visual features. They are followed by LBP-based global texture features, which in turn outperform the mid-level features such as harmony, dynamism, *etc.* The DAISY feature does not provide a MiAP performance as good as SIFT features, but it uses a shorter descriptor length and operates 3 times faster [17]. Moreover, four multi-scale color LBP features perform better than the original LBP as they possess the enhanced photometric invariance property and discriminative power [19]. The mid-level features (such as dynamism and harmony) yield the lowest performance. The reason could be that these mid-level features are global ones and generally of very low dimension, *e.g.* only 1 value for harmony and dynamism, respectively. Therefore, they may not capture sufficient local visual information as compared to local features, *e.g.*, C-SIFT with BoVW modelling, which are much more comprehensive.

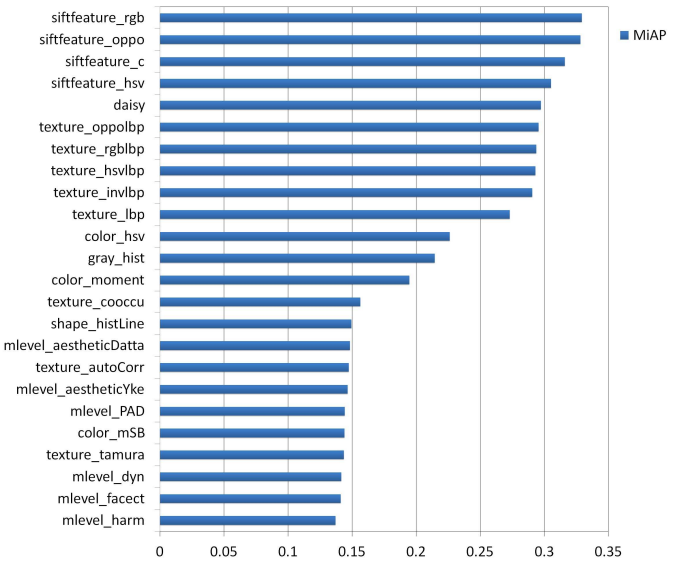


Figure 5: The MiAP performance of each visual features on the validation set.

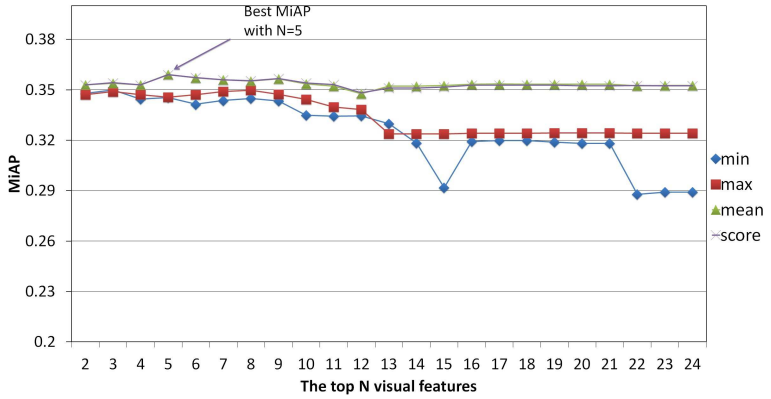
We performed the SWLF scheme for fusing visual features, and found that fusing the top 5 features yield the best MiAP (35.89%) on the validation set, as shown in Figure 6 (a). The results indicated that the weighted score and mean rules through SWLF outperforms the other two fusion rules, namely min and max, and the MiAP performance is increased by 3% using the weighted score-based SWLF scheme compared to 32.9% achieved by the best single visual feature (RGB-SIFT). As a result, the visual model, which we submitted to the photo annotation task at ImageCLEF 2011, performed the fusion of the top five best visual features using the score-based SWLF scheme. As shown in Figure 6 (b), the fused experts proved to have a very good generalization skill on the test set. It can be seen that the weighted score and mean fusion methods perform better than the others, and the best fused experts on the validation set, which combine the top 5 features, achieved a MiAP of 35.54% on the test set, in comparison with a MiAP of 35.89% achieved by the same fused experts on the validation set.

As it can be seen from Figure 6, the performance by score-based SWLF is not that different from the performance by mean-based SWLF even though the former performs slightly better than the latter, especially on the test set. The reason is that the weights computed by SWLF for all the experts are not that different, even roughly the same for the 10 best visual features,

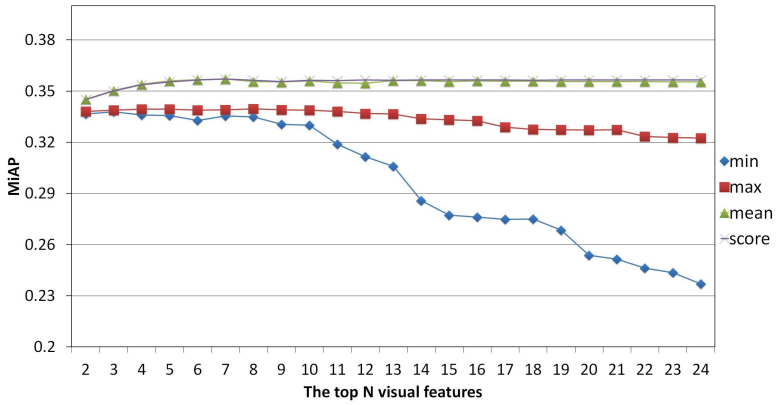
e.g., SIFT-like and LBP-like features. As it can be seen in figure 5, the first 10 features, *e.g.* SIFT-like features and LBP-like features, display roughly similar MiAP ($\cong 0.3$) whereas the last 11 features, *e.g.*, mainly all the mid-level features, also have a similar but lower MiAP ($\cong 0.15$). In the SWLF algorithm, each expert is weighted by its normalized iAP. This eventually results in roughly the same weights for the 10 first experts, and a range of weights which is not that big when N goes beyond 10, especially after weight normalization. Now, when these experts are fused by SWLF using the number of experts N increased from 2 to 24, the performance difference in terms of MiAP between the mean and score fusion rules through SWLF are really hardly noticeable at the scale of accuracy in Figure 6 until N reaches 10, this difference becomes more apparent when the number of experts N fused by SWLF goes beyond 10. A higher difference between these two fusion rules would certainly be observed if the performance of the features was very different.

5.3. Experimental results on the textual modality

The textual features were described in Section 2. As follows the study by Bradley *et al.* [52] who pointed out that the measurement of dominance values is not as stable as those of valence and arousal, we excluded from our study the textual features making use of dominance values, namely *txtf_1034pvad* and *txtf_1034wvad*, and only investigated the remaining ten variants of HTC summarized in Table 1 in our participation to the ImageCLEF 2011 photo annotation task. All these 10 types of textual features were first extracted from the training set and used to train different SVMs using two different kernels, namely χ^2 and RBF. It is interesting to test and evaluate these two kernels since RBF kernel simply assumes a radial distance in a higher dimensional space where each sample is nonlinearly mapped and is a reasonable first choice, whereas χ^2 kernel measures distributional similarities between two histograms of occurrence frequencies. On the other side, HTC feature has the form of a histogram but differs from traditional histogram as defined statistics and probability. Indeed, we remind that the value for each bin in HTC, *i.e.* a textual concept of a given vocabulary, instead of counting term frequency, accumulates the relatedness of that concept towards each textual label associated to an image. On the validation set, it turns out that the RBF kernel outperforms the χ^2 kernel in terms of MiAP by 15 points. Figure 7 shows the MiAP performance of different types of textual features using the RBF kernel.



(a)



(b)

Figure 6: The MiAP performance of different fusion methods based on SWLF scheme using the visual features on the validation set (a) and test set (b). As required by SWLF, the features are first sorted by descending order in terms of iAP of their corresponding experts. Then, the number of fused features N is increased from 1 to 24 (total number of visual features).

As we can see from Figure 7, the best MiAP, close to 0.3, was achieved by the HTC variant *txtf_99ps* using the dictionary *D_99*, the *path* distance and the *sum* accumulator while the textual features making use of affect coordinates in terms of valence and arousal achieved the lowest MiAP. After our submission, we carried out additional experiments using the textual features *txtf_1034wvad* and *txtf_1034pvad* that contains dominance informa-

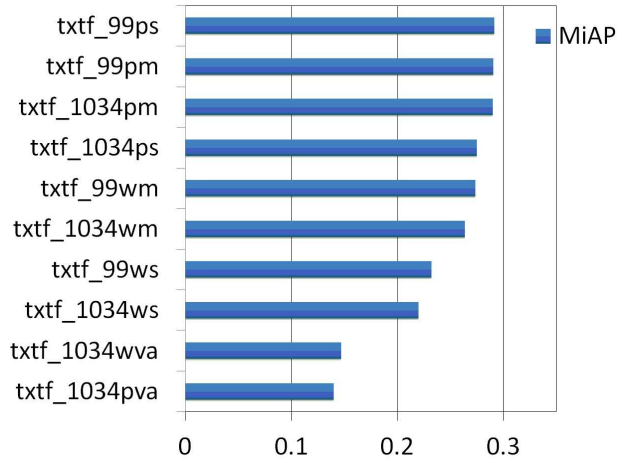


Figure 7: The MiAP performance of textual features on the validation set.

tion along with valence and arousal values. We discovered that the addition of the dominance information improves by 3 or 4 percent the overall performances achieved by the textual features without dominance values, namely *txtf_1034wva* and *txtf_1034pva*. However, the performances achieved by *txtf_1034wvad* and *txtf_1034pvad* stay still quite far away from the MiAP displayed by the best textual feature, *e.g.*, *txtf_99ps*.

The performance of the textual features is thus lower than that displayed by visual features as shown in Figure 5. However, textual features behave much differently from the visual ones. The tags associated with images may provide valuable cues to the visual semantic content so as to correct the misclassification by the visual features, and their joint use should lead to the improvement of predictions. To illustrate this, we show in Figure 8 several images from the validation set that were misclassified using the visual modalities, but correctly classified by the textual ones. In the first example, the image is taken inside the airplane, which is unusual in the training set and makes the visual modality fail to detect the "airplane" concept. However, the associated text contains a "plane" tag, and our textual feature HTC successfully captures this cue and shows a high value for the bin associated with the concept "airplane", thereby facilitating a correct prediction by a textual classifier.

We also applied the SWLF scheme to fuse textual features. The results as shown in Figure 9 (a) indicate that the combination of the top 5 best fea-

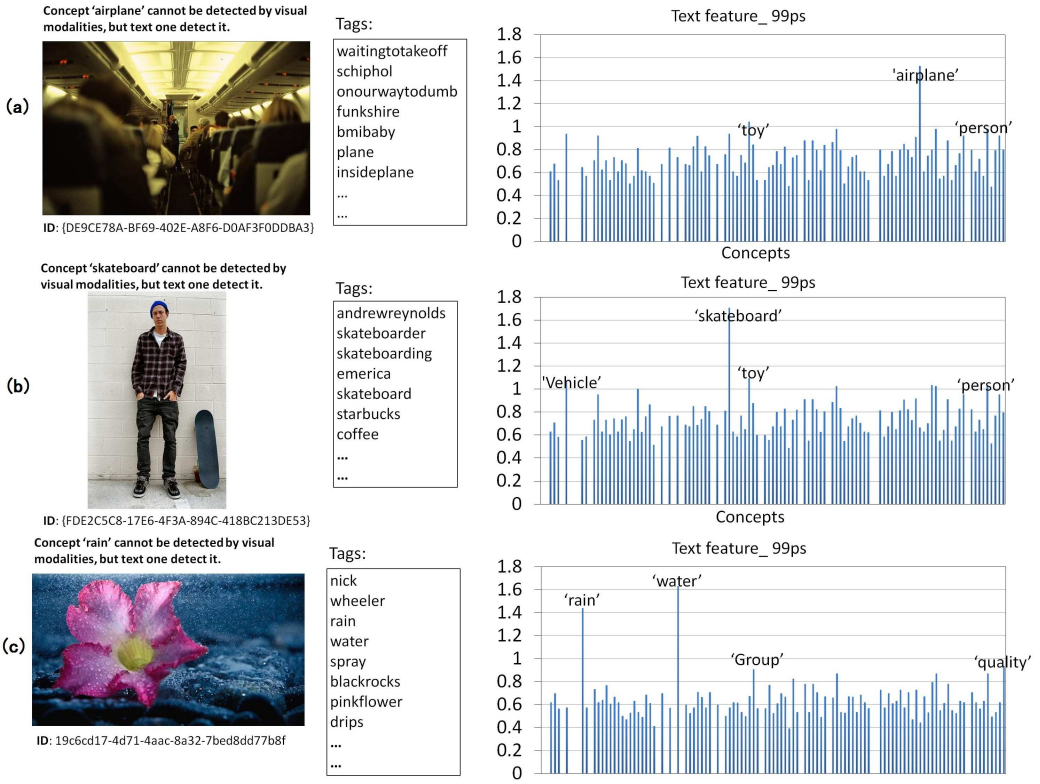
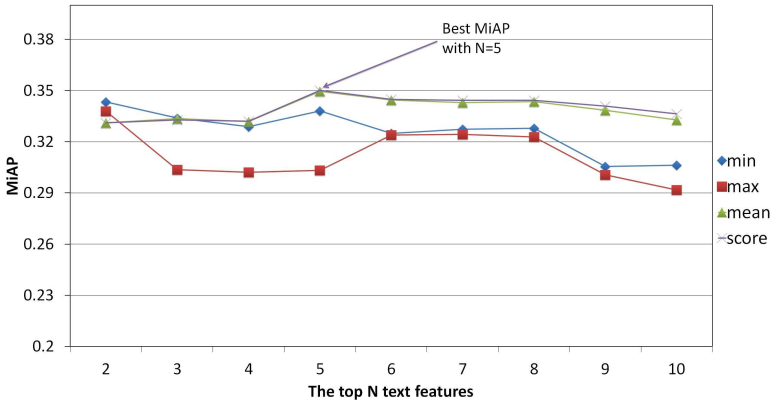


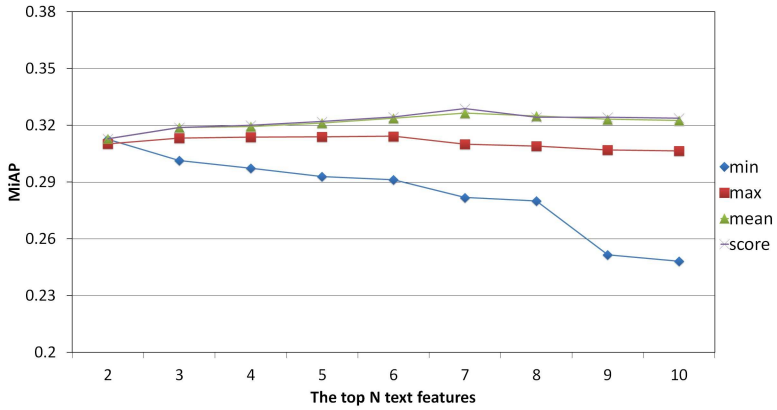
Figure 8: The left column shows some visual concepts which can be hardly predicted by the visual modalities, but can be predicted by the textual ones. The center column shows the raw text tags associated with the photos; the right column shows the built textual feature vector HTC using *txtf_99ps*. As shown in (b), the skateboard is imaged in a unusual viewing angle with only 6 positive training samples, leading to the misclassification of classifiers using only the visual modality. On the other hand, the term "skateboard" appears in the user tags, which is successfully captured by the textual feature vector HTC, thereby facilitating the correct prediction of the classifiers using the textual modality.

tures yield the best MiAP value, and the weighted score-based SWLF scheme outperforms the other fusion rules, and achieves a MiAP of 35.01% which improves by 6 points the MiAP of 29.1% achieved by the best single textual feature (*txtf_99ps*) on the validation set. As a result, we implemented our text-based prediction model using the weighted score-based SWLF scheme to fuse the top 5 best textual features. As shown in Figure 9 (b), the fused experts using the top 5 features achieve a MiAP of 32.12% on the test set. It thus displays a very good generalization skill when this last figure is compared with the MiAP of 35.01% achieved by the same fused experts on the validation set. Again, we also discovered that the score and mean-based SWLFs perform better than the others.

To compare HTC features with the popular BoW approaches, we implemented the typical TF and TF-IDF models, which are based on word counts or term frequency. Figure 10 compares the MiAP performances of TF, TF-IDF and HTC features on the test set. It further suggests that HTC features, in accounting for the relatedness of concepts and smoothing the histograms of textual concepts, prove to be effective, in particular when dealing with sparse user tags. We also compared our textual feature HTC with an extended BoW approach, namely the Latent Dirichlet allocation (LDA) method [31], which views a document as a mixture of topics with a dirichlet distribution prior. This topic model can help gain insight into the latent topics within the text and enables the reduction of the high dimensionality of the feature space. The LDA approach with 64 topics achieved a MiAP of 13.2% as indicated by the red line in Figure 10. The results indicate that the proposed HTC outperforms LDA method with almost 15 points in terms of MiAP. The main reason is that image captions are generally sparse texts, only having on average 8.7 tags per image for example in MIR FLICKR image collection. Thus, they may not provide enough text content to correctly train the LDA topic model.



(a)



(b)

Figure 9: The MiAP performance of different fusion methods based on the SWLF scheme using textual features on the validation set (a) and test set (b). As required by SWLF, the features are first sorted by descending order in terms of iAP of their corresponding experts. Then, the number of fused features N is increased from 1 to 10 (total number of textual features).

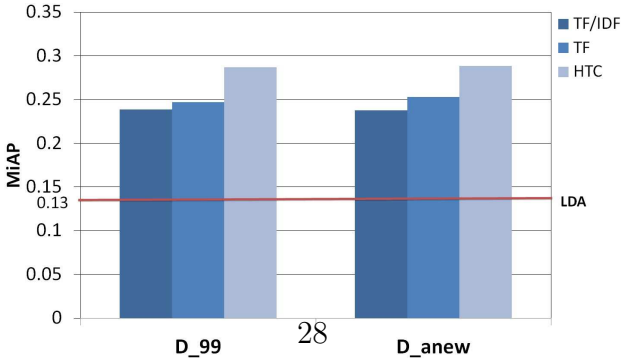


Figure 10: The MiAP performance of different textual approaches on the test set.

5.4. Experimental results on fusing textual and visual features

We also implemented our multimodal approach using the SWLF scheme as described in Section 4 to fuse the visual and textual features. Recall that we implemented 24 visual features and 10 textual features. Figure 11 shows the MiAP performance achieved by the different variants of the SWLF scheme on the validation and test sets as N, the number of features to be fused, is increased from 1 to 34.

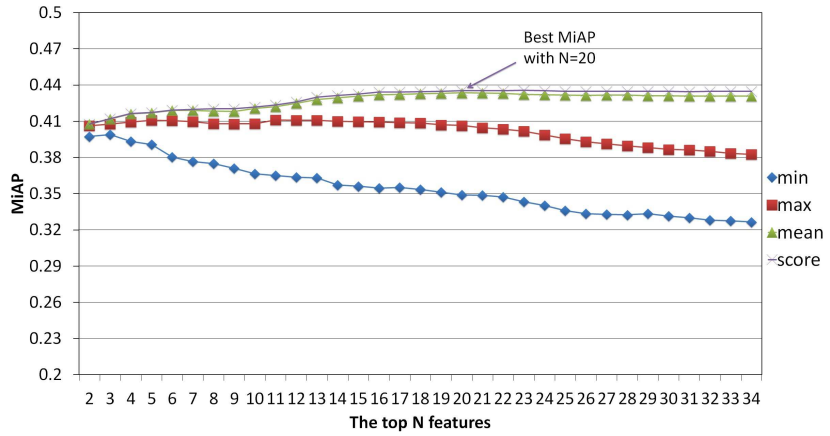
As we can see from Figure 11 (a), the MiAP performance of max and min-based SWLF schemes tend to decrease when the number of features, N, is successively increased from 1 to 34, while the performance of weighted score and mean-based SWLF schemes keep increasing until N reaches 20 and then stays stable. The weighted score-based SWLF scheme performs slightly better than the mean-based SWLF scheme. The weighted score-based SWLF scheme using $N = 20$ displays a MiAP of 43.54% and increases thus by 9 points the MiAP compared to 34.1% achieved by the best textual prediction model, and by 7.6 points compared to 35.9% achieved by the best visual prediction model. These results demonstrate that the weighted score-based SWLF scheme performs consistently much better than the max and min-based fusion rules and leads to a slightly better performance than the mean-based variant. Figure 11 (b) shows the performance on the test set of the fused experts combining textual and visual features. We can see from that figure that the results are very similar to the one achieved on the validation set, which prove that fused experts present a very good generalization skill.

5.5. Experimental results on affect related features

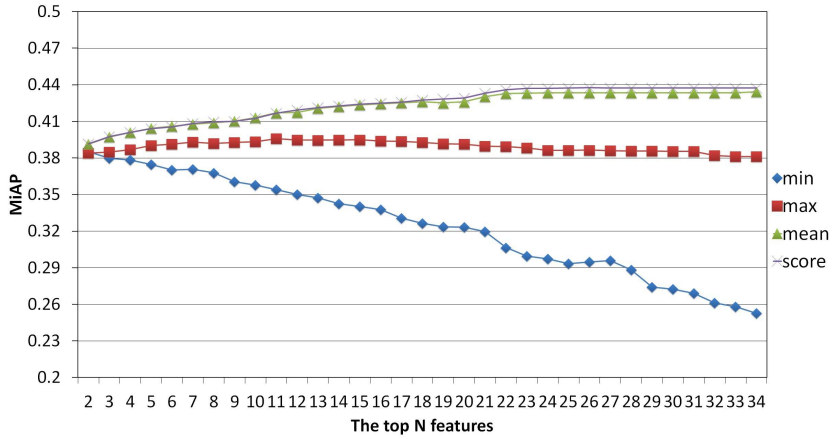
The photo annotation task at ImageCLEF 2011 featured 9 affective concepts, namely "active", "euphoric", "funny", "happy", "calm", "inactive", "melancholic", "unpleasant" and "scary". For this purpose, we investigated several affect related descriptors, both in textual and visual modalities, in order to evaluate their usefulness regarding to these affective concepts.

5.5.1. On the usefulness of affective textual features

Using the textual features on the ANEW dictionary, we implemented in Section 2 the coordinates of an image caption in terms of valence and arousal, namely $f_{valence}$ and $f_{arousal}$ using Equation (1) and (2). Figure 7 clearly shows that the performance of these two affective textual features $f_{valence}$ and $f_{arousal}$ is lower than that of the other HTC-based features in terms of MiAP (averaged on all concepts), one may think that these features



(a)



(b)

Figure 11: The MiAP performance of different fusion methods based on SWLF scheme using visual and textual features on the validation set (a) and on the test set (b). As required by SWLF, the features are first sorted by descending order in terms of iAP of their corresponding experts. Then, the number of fused features N is increased from 1 to 34 (total number of features).

could be effective for the prediction of affective concepts. As these affective textual features deliver only two values, we evaluated on the validation set the effectiveness of $f_{valence}$ in discriminating the concept *happy* from the concept *unpleasant* at the valence axis and the one of $f_{arousal}$ in distinguishing the concepts *active* and *inactive* at the arousal axis, using the training set for

learning.

Table 3 and 4 shows the confusion matrix of the two features. These results, unfortunately close to random, show that the prediction of emotional concepts is extremely challenging. They can be explained by several reasons: the lack of sufficient training data, the subjective nature underlying these affective concepts, the empiric ratings of the ANEW words, *etc.*.

Table 3: The confusion matrix of $f_{valence}$ for *happy* and *unpleasant* categories.

Actual \ Predicted	happy	unpleasant
happy	50.60	49.40
unpleasant	56.47	43.53

Table 4: The confusion matrix of $f_{arousal}$ for *active* and *inactive* categories.

Actual \ Predicted	active	inactive
active	46.40	53.60
inactive	47.26	52.74

5.5.2. The impact of the dictionary underlying HTC: D_99 versus D_Anew

We also compared the performance of the HTC features using different dictionaries, respectively *D_99* and *D_Anew*, shown in Figure 12. Recall that *D_99* is composed of the 99 visual concepts in the photo annotation task at ImageCLEF 2011 while *D_Anew* is composed of 1034 English words for which the affective coordinates were defined in ANEW study. The *D_99* fully describes the semantic space spanned by the 99 visual concepts while *D_Anew* should have some impact on the recognition of the 9 affective concepts. Figure 12 shows the iAP performance of the feature *txtf_99ps* using *D_99* and the feature *txtf_1034ps* using *D_Anew* on each of 99 visual concepts. From the figure, we can see that *txtf_99ps* performs slightly better than *txtf_1034ps* (31.38% vs. 30.38%), but *txtf_1034ps* outperforms *txtf_99ps* on the concepts related to emotions, *e.g.*, "happy", "funny", "euphoric", "scary", "unpleasant" and "melancholic". The reason could lie in the fact that *D_Anew* is built

with the ANEW dictionary that may contain more words describing the emotions. Thus, the emotional tags are more likely to be better described with *D_Anew* than *D_99*. This intuition is further confirmed by the fact that, for some concepts, *e.g.* "skateboard", which is not included in *D_Anew*, the HTC features built with *D_99* achieve much higher iAP than the ones with *D_Anew* (54.7% vs. 0.23%).

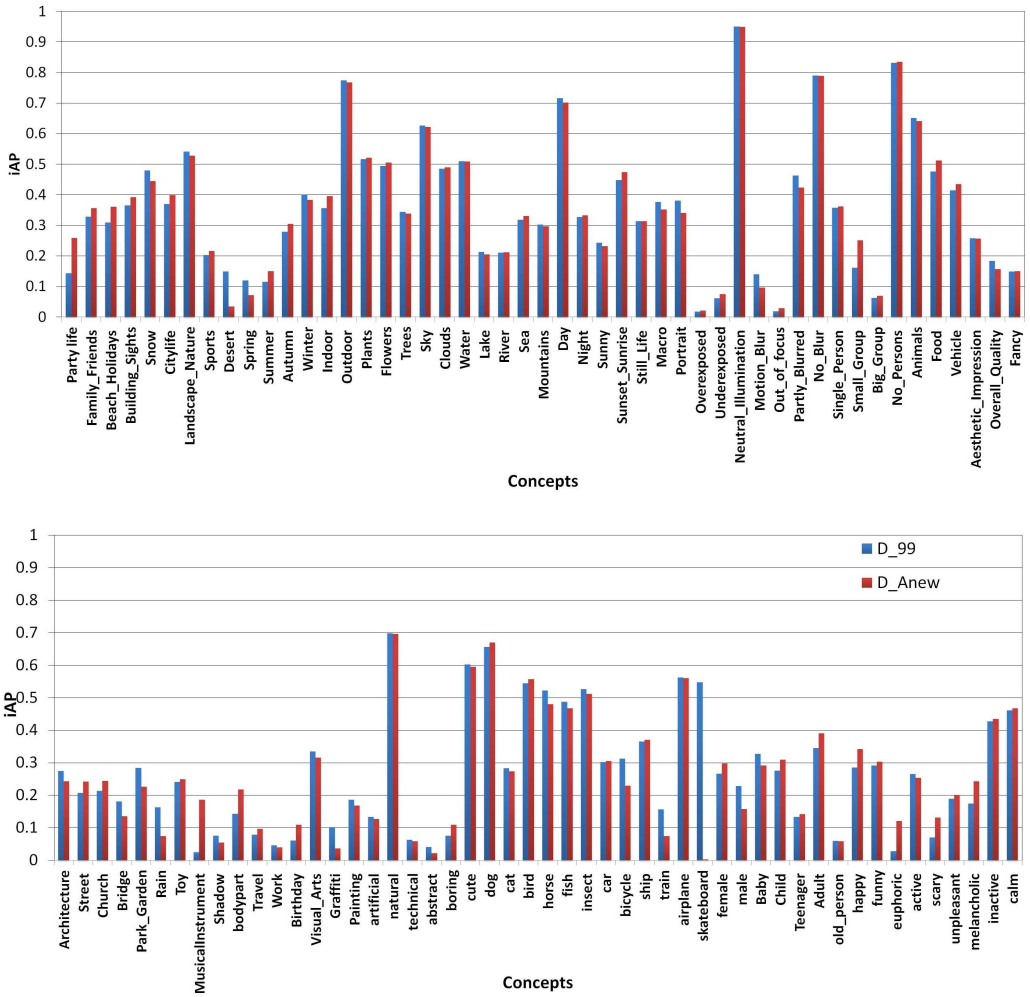


Figure 12: The average precision performance of the textual features *txtf_99ps* vs. *txtf_1034ps*, respectively built with the dictionaries *D_99* and *D_Anew*.

5.5.3. On the usefulness of affective visual features

We further evaluated the set of mid-level visual features, including harmony, dynamism, aesthetic quality, *etc.*, for their usefulness in predicting the nine affect concepts. Figure 13 shows the prediction capability of each mid-level feature on the nine emotion classes in comparison with the one obtained by the best local RGB-SIFT.

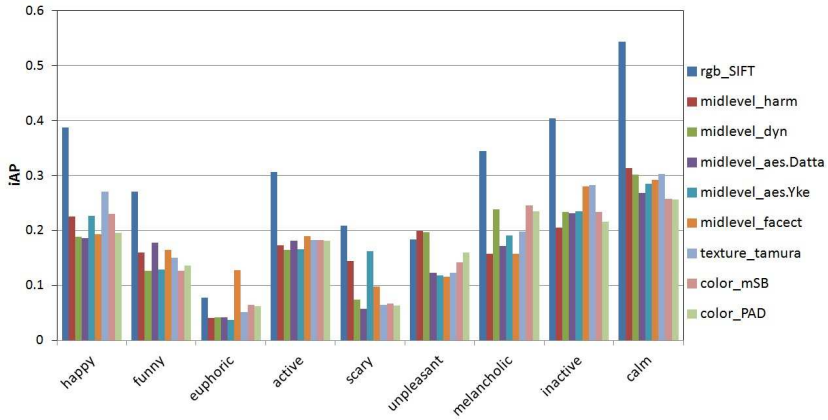


Figure 13: The iAP performance of affect related visual features on the 9 sentimental concepts on the validation set compared to the best single visual feature RGB-SIFT.

Our intuition was that adding mid-level affect-related features would help bridge the semantic gap between low-level features and very high-level semantic concepts such as emotions. However, these results show that low-level RGB-SIFT outperforms all mid-level features for almost all the nine affect concepts. The reason may lie in the fact that these mid-level features are global features and do not capture sufficient visual content as compared to the local descriptors. This also further confirms that machine-based prediction of affective concepts is extremely challenging. In particular, the subjective nature of these concepts makes it hard to have homogeneous manually labeled data. Moreover, the lack of sufficient training data further complicates this task.

5.6. Experimental results of the ImageCLEF 2011 photo annotation challenge

We submitted 5 runs to the ImageCLEF 2011 photo annotation challenge (2 textual prediction models, 1 visual prediction model and 2 multimodal prediction models). All runs were evaluated on the test set composed of 10000

images. They were learnt by the weighted score-based SWLF on the training and validation sets using the features described in the previous sections, including 10 textual ones (using user tags) and 24 visual ones. The two textual prediction models made use of only textual features extracted from the user tags associated with an input image for predicting the visual concepts within it. The visual prediction model made use of only visual features while the two multimodal prediction models made joint use of the textual and visual features. We did not use the EXIF meda data provided for the photos.

1. **textual_model_1**: the combination of the top 4 features among the 10 textual features for each concept based on the weighted score SWFL scheme.
2. **multimodal_model_2**: the combination of the top 21 features among 34 visual and textual features for each concept based on the weighted score SWFL scheme.
3. **textual_model_3**: the combination of the top 5 features among the 10 textual features for each concept based on the weighted score SWFL scheme.
4. **visual_model_4**: the combination of the top 5 features among the 24 visual features for each concept based on the weighted score SWFL scheme.
5. **multimodal_model_5**: the combination of the top 22 features among the 34 visual and textual features for each concept based on the weighted score SWFL scheme.

Table 5: The results of our submitted runs.

Submitted runs	MiAP(%)	F-Ex(%)	SR-Precision(%)
textual_model_1	31.76	43.17	67.49
multimodal_model_2	42.96	57.57	71.74
textual_model_3	32.12	40.97	67.57
visual_model_4	35.54	53.94	72.50
multimodal_model_5	**43.69**	**56.69**	**71.82**
Best MiAP: TUBFI	44.34	56.59	55.86

Thanks to the combination of the textual and visual features using our weighted score-based SWFL scheme, our 5^{th} multimodal run achieved a MiAP of 43.69% which was ranked the 2^{nd} performance out of 79 runs on the MiAP evaluation, as shown in Table 5. Indeed, our best visual model with 35.5% was awarded the 5^{th} in comparison to the best performance of 38.8% in visual configuration. Our best textual model with 32.1% was ranked the 4^{th} performance while the best performance of textual modality was 34.6 %. Our weighted score-based SWLF fusion method again demonstrated its effectiveness, displaying a MiAP of 43.69% which improves the MiAP of 35.54% of our visual prediction model by roughly 8% and even by 11% the MiAP of 32.12% of our best textual prediction model.

5.7. Discussion

In this work, we designed a novel textual feature, namely HTC, for the relatedness of textual concepts and proposed a Selective Weighted Late Fusion (SWLF) scheme to best select and fuse the features for the purpose of VCR. While one could want to know the real role the proposed HTCs played in our submissions in comparison with the visual features, another question which naturally arises is how the size of the validation set impacts the overall performance on unseen test data, as SWLF requires a validation set to best select and fuse an ensemble of experts. We discuss these two questions in this subsection and start with the study of the impact of the size of the validation set on the overall performance.

5.7.1. Impact of the size of the validation set on the generalization skill of the fused expert through SWLF

In our experimental setup, defined in Section 5.1, the initial training dataset was divided into two roughly equal parts: a training set and a validation set. The SWLF uses the training set to train an ensemble of experts (classifiers), one for each visual concept and each type of features. It then selects and combines the best experts while optimizing the overall MiAP on the validation set.

The first question concerns the generalization skill of a fused expert through SWLF on unseen data. In Section 5.2, 5.3 and 5.4, we already depicted the good generalization skill of the fused experts through the weighted score-based SWLF, on test dataset. Table 6 further highlights such a behaviour of the fused experts in displaying their MiAP performance both on the validation and test dataset. The prediction models in bold correspond

to the best prediction model learnt through SWLF on the validation set. It can be seen that the best fused experts learnt on the validation set keeps a quite good generalization skill as the performance only drops slightly on the test set. In our submission, we anticipated this performance drop in particular for multimodal prediction models. Instead of submitting the best multimodal model on the validation set which combines the best 20 features, we submitted two multimodal runs, namely multimodal_model_4 and multimodal_model_5, making use of 21 and 22 best features, respectively. Surprisingly enough, our best multimodal run, multimodal_model_5, which was ranked the second best MiAP performance out of 79 runs, proves to perform slightly better on the test set than on the validation set.

Table 6: MiAP performance comparison of the fused experts learnt through the weighted score-based SWLF on the validation set versus the test set. The prediction models in bold correspond to the best fused experts learnt through weighted score-based SWLF on the validation set.

Prediction model	Nb of fused experts N	Validation set	Test set
textual_model_1	4	33.21	31.76
textual_model_3	5	35.01	32.12
visual_model_4	5	35.89	35.54
multimodal_model	20	43.54	42.71
multimodal_model_2	21	43.52	42.96
multimodal_model_5	22	43.53	43.69

The second question is how the size of the validation set impacts the generalization ability of a fused expert learnt through SWLF. For this purpose, we evaluated, as shown in Figure 14, the performance of the fused multimodal experts learnt through the score-weighted SWLF on the validation set, by varying the size of that validation set. The results on the test set were achieved by varying the size of the validation set from 20% to 100% of the size of the original validation set, *i.e.* 3995 images as specified in Section 5.1, while keeping the training set unchanged. The x axis displays the number of fused experts while the y axis gives the MiAP performance. The curves in different colors plot the MiAP performance using different size of the validation set. From this figure, we can see that the SWLF performance keeps increasing with the size of the validation set, and the improvement

becomes slight from 40% of the size of the original validation set. Given the size of a validation set, the fused expert displays a similar behaviour: the performance increased quickly when N varies from 1 to 20, then it subsequently remains stable.

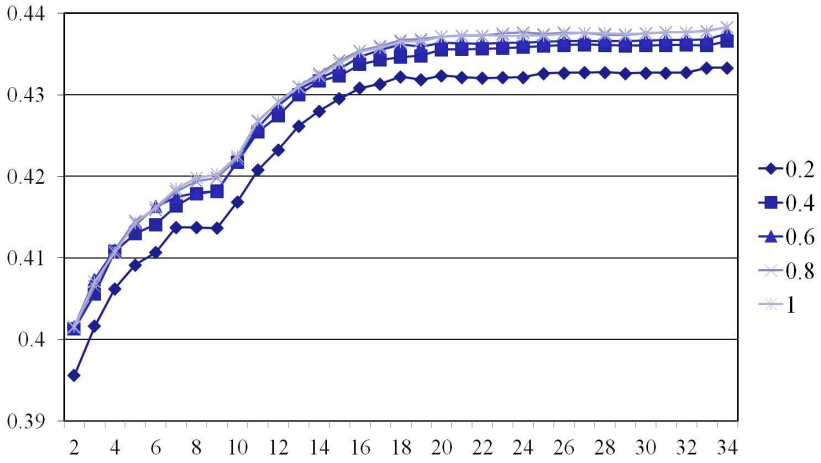


Figure 14: The MiAP performance on the test dataset of the fused experts through SWLF when varying the size of the validation dataset from 20% to 100% of the size of the original validation set.

5.7.2. Role of HTC features on the MiAP performance of the proposed prediction models

In Section 5.3, we have already highlighted the usefulness of the proposed textual HTC features to predict some visual concepts with a very small number of training samples, *e.g.*, "airplane","skateboard","rain". In this section, we further investigate the contributions of the proposed HTC features to the overall performance achieved at the photo annotation task within ImageCLEF 2011 challenge. Our best run, the prediction model *multimodal_model_5*, achieved the best iAP performance on 13 visual concepts out of 99 [9]. Figure 15 presents the occurrence frequency of our textual features within the top 5 feature selected by the SWLF for the 13 concepts. We can see that our textual features greatly contribute to the performance of our approaches and improve our final ranking in the challenge.

To further emphasize the contribution of the proposed textual features, we present in Table 7 the top 22 features selected by *multimodal_model_5* for those 13 concepts. As we can see, there are 7 concepts in which the HTC

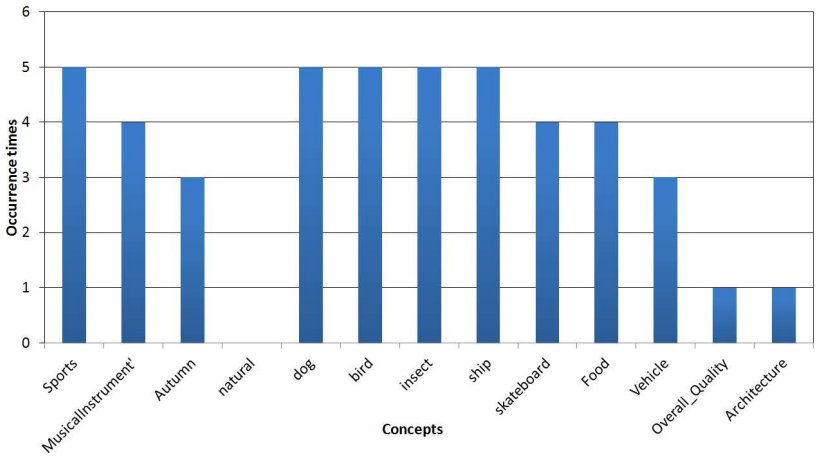
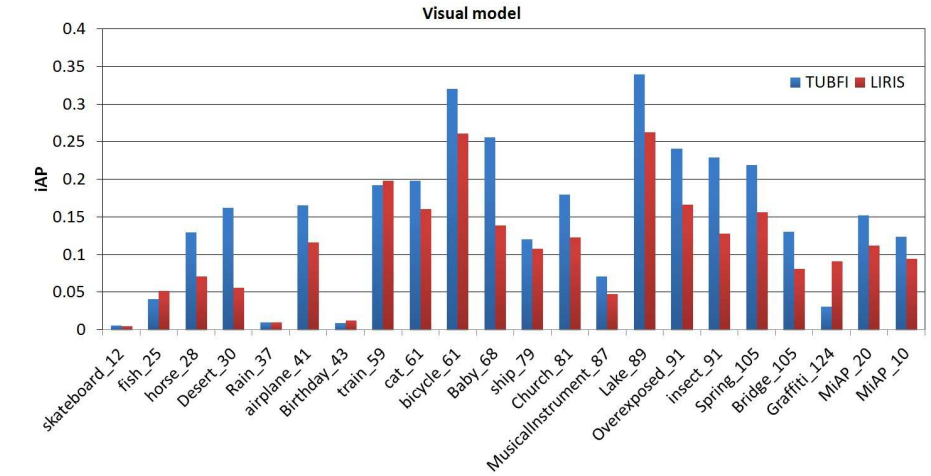


Figure 15: The occurrence frequency of textual features within the top 5 features selected by the SWLF for the 13 concepts for which we achieved the best iAP values in the ImageCLEF 2011 Photo annotation task.

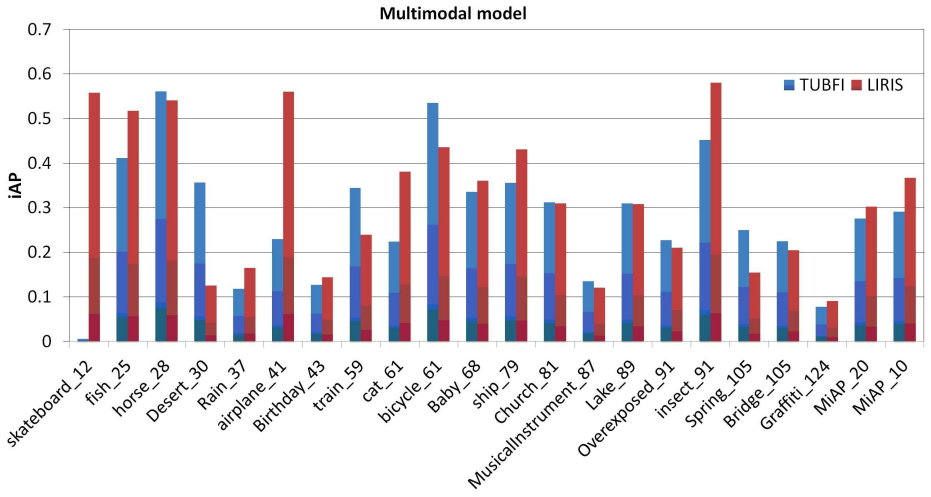
features achieved the best iAP performance, while the most powerful visual features were the local descriptors and multi-scale color LBP operators.

Table 7: The top 22 features selected by SWLF within *multimodal_model_5* for the 13 concepts for which we achieved the best iAP values in the ImageCLEF 2011 photo annotation task. The numbers in each column give the rank of each feature with the iAP value in brackets.

	7 Sports	56 Musical Instrument	11 Autumn	66 natural	71 dog	73 bird	76 insect	79 ship	82 skateboard	44 Food	45 Vehicle	47 Overall Quality	49 Architecture
grey_hist	21(4.15)	16(6.18)	12(12.8)	16(64.7)	19(14.2)	20(6.62)	19(11.8)			20(15.3)	19(21.1)	19(26.4)	16(26.1)
color_hsv		10(11.5)	22(4.24)	11(67)	20(14.2)	13(14.1)	18(11.9)	6(10.6)		18(20)	20(16)	9(32.4)	15(26.7)
texture_lbp	3(12.8)	12(8.06)	8(15.7)	10(67.2)	17(16.6)	18(7.47)	17(12.8)	18(5.76)	6(3.24)	19(19.1)	17(29.2)	2(36.7)	3(36.5)
texture_hsvlbp	4(12.7)	4(15.2)	11(13.5)	9(70)	14(20.8)	12(14.2)	11(16.5)	16(7.14)	7(1.36)	15(28)	16(30.1)	3(36)	5(35.2)
texture_invlbp	11(8.7)	1(18.8)	9(14.4)	7(70.9)	16(18.1)	17(9.06)	14(16)	14(7.47)	4(5.73)	16(27.4)	15(30.8)	4(35.9)	7(35)
texture_rgblbp	6(12.6)	17(5.81)	10(14.4)	8(70.5)	18(15.9)	14(13.6)	16(14.9)	17(7.01)	5(3.88)	14(28.1)	12(33.3)	1(37.1)	8(34.8)
texture_oppolbp	5(12.7)	6(13.6)	6(19.5)	6(71)	15(19.7)		9(17.1)	4(14)	22(0.26)	13(28.7)	13(32.4)	8(34.4)	9(34.3)
color_moment		18(5.13)		13(65.3)		21(5.64)	21(5.16)	7(10.6)	14(0.586)	21(9.06)			
color_mSB		22(3.54)				22(4.7)							22(22.4)
color_PAD		9(11.8)		18(63.8)	21(12.4)				8(1.13)				
texture_tamura	20(4.25)					19(7.43)					22(14.5)		
texture_cooccu		21(3.73)	21(5.61)	20(63.4)	22(7.39)			21(3.33)			21(14.8)		
texture_autoCorr		19(4.97)							17(0.485)			14(29.1)	
shape_histLine													
mlevel_harm									19(0.428)	22(8.77)			
mlevel_dyn												17(28)	
mlevel_aestheticDatta	19(4.42)												20(23.2)
mlevel_aestheticYke			18(6.58)									15(28.7)	
mlevel_facect								22(3.21)				18(27.3)	
text_feature99ps	9(9.89)		15(11.2)	14(64.8)	1(69.7)	3(52.9)	5(40.6)	12(8.41)	2(9.29)	9(36.1)	9(34.6)		14(28.6)
text_feature99pm	8(10.1)		14(11.2)		2(67.9)	2(57.5)	6(39.8)	1(15.8)	1(9.31)	12(32.6)	6(35.2)	21(25.8)	21(22.7)
text_feature99ws	2(13.8)		20(5.76)		7(35)	7(29.2)	7(31.5)	19(5.49)	11(0.861)	6(36.8)	11(33.4)	22(25.5)	17(24.5)
text_feature99wm	1(16.6)	14(7.57)	19(6.13)	19(63.6)	6(40.8)	5(46.6)	3(45.8)	9(9.38)	3(6.66)	4(39.3)	4(36.8)	13(30.3)	11(30.5)
text_feature1034ps	12(8.29)	15(6.88)	13(12.1)	15(64.7)	4(66.1)	4(47)	4(44.3)	3(14.3)	16(0.513)	3(40.1)	10(33.6)		12(30.4)
text_feature1034pm	10(9.64)	13(7.93)	3(24.3)	17(64.2)	3(67.6)	1(58)	1(49.8)	5(10.7)	12(0.826)	5(37.2)	14(31.6)	10(31.5)	19(24)
text_feature1034ws	13(7.17)		16(8.16)	12(66.6)	9(30.3)	8(24.2)	8(31)	20(3.56)	9(1.02)	2(40.9)	18(28.8)		18(24.1)
text_feature1034wm	7(10.9)	5(14.6)	17(7.17)	21(62.8)	5(42)	6(41.7)	2(48.4)	10(8.58)	10(0.963)	1(42.7)	8(34.8)	16(28.2)	13(30.1)
text_feature1034pva	22(3.62)			22(62.3)					13(0.81)				
text_feature1034wva		20(4.07)					22(4.56)					20(25.9)	
siftfeature_c	16(6.1)	7(13)	4(23.7)	3(71.9)	12(27.7)	15(12.1)	15(15.9)	11(8.46)		10(36)	5(36.6)	6(35.1)	6(35.1)
siftfeature_rgb	17(5.13)	2(17.5)	1(26.3)	2(72)	11(29.5)	9(17.4)	13(16.2)	8(10.4)	18(0.432)	8(36.2)	2(40.9)	7(35)	1(38.3)
siftfeature_hsv	18(4.88)	3(15.3)	5(22)	4(71.2)	10(29.7)	16(11.9)	10(16.6)	13(7.66)	20(0.419)	11(35)	7(35.1)	11(31.1)	4(35.5)
siftfeature_oppo	15(6.17)	11(10.4)	2(24.5)	1(72.2)	8(32.7)	10(16.3)	12(16.3)	2(15.1)	21(0.398)	7(36.6)	1(42.1)	5(35.8)	2(37.6)
daisy	14(6.29)	8(12.6)	7(16.4)	5(71.1)	13(24.6)	11(15.3)	20(8.77)	15(7.41)	15(0.584)	17(23.4)	3(39.5)	12(30.9)	10(32.6)



(a)



(b)

Figure 16: The iAP performance of our visual and multimodal prediction models, namely visual_model_4 and multimodal_model_5, compared to the best TUBFI's runs on 20 concepts having the smallest set of training samples. These concepts were selected according to the size of their positive training samples in an ascending order. "skateboard_12" denotes that 12 training samples are provided for that concept. (a) compares the iAP performance by TUBFI with our submitted visual model, visual_model_4. The TUBFI's best visual model achieves a MiAP of 15.23% and 12.32% on the top 20 and 10 concepts having the smallest number of positive samples, respectively, in comparison with 11.2% and 9.4% by our run using visual_model_4. (b) shows the performance of the submitted multimodal models on these concepts: our multimodal_model_5 run achieves a MiAP of 30.1% and 36.65% on the top 20 and 10 concepts having the smallest set of positive samples, respectively, in comparison with 27.58% and 29.11% by TUBFI's best multimodal run.

In Section 5.3, we highlighted the interest of the proposed HTC features which prove to be particularly useful when the training data is small for a given visual concept. They provide significant complementary information to the visual features. Figure 16 further spotlights such behaviour in plotting the performance of our textual and visual features on the top 20 concepts having the smallest training set, and compares our results with those achieved by TUBFI's multimodal run which, with a MiAP of 44.34%, was ranked the best performance in the challenge. It shows that the MiAP of our prediction model, multimodal_model_5, outperforms TUBFI's by about 8 % in the first 10 concepts. it can be seen that our multimodal prediction model significantly outperforms the TUBFI's best run on the concepts "airplane" and "skateboard". Indeed, the number of training samples for these concepts are only 41 and 12 respectively, thus making it extremely difficult to correctly train classifiers and to accurately predict those concepts if only visual features were used. TUBFI's multimodal run achieved a iAP of 22.93% and 0.56% for "airplane" and "skateboard" concepts. In contrast, our textual features significantly improve the performance of our visual classifiers with regard to these cases. Using our multimodal prediction model learnt through weighted score-base SWLF, the proposed textual features enhance the MiAP performance of the visual configuration from 11.62% to 56.01% for "airplane" and from 0.45% to 55.79% for "skateboard".

6. Conclusion

In this paper, we investigated a multimodal approach for the purpose of VCR. Firstly, we proposed a novel textual descriptor, namely the Histogram of Textual Concepts (HTC), which relies on the semantic similarity between the user tags and a concept dictionary. We also evaluated a set of mid-level visual features, aiming at characterizing the harmony, dynamism and aesthetic quality of visual content, in relationship with affective concepts. Finally, a novel selective weighted late fusion (SWLF) scheme was also introduced which iteratively selects the best features and weights the corresponding scores for each concept at hand to be classified.

The extensive experiments were conducted on a subset of the MIR FLICKR collection used in the photo annotation task in the ImageCLEF 2011 challenge, where our best multimodal prediction model achieved a MiAP of 43.69% and ranked the 2^{nd} best performance out of 79 runs. From a comprehensive analysis of the experimental results, we can conclude the following:

(i) the proposed textual HTC features greatly improve the performance of visual classifiers, especially when the training set of a given concept is small; (ii) the recognition of emotion related concepts is still extremely challenging and the set of mid-level affect-related features that we implemented did not help bridge the semantic gap between affective concepts and low-level features, and it surprisingly turn out that the low-level features perform better than mid-level ones. The reason may lie in the fact that these mid-level features are global features and don't capture sufficient visual content as compared to local descriptors; (iii) the fused experts through weighted score-based SWLF, which best selects and combines an expert ensemble while optimizing an overall performance metrics, *e.g.*, MiAP in this work, display a very good generalization skill on unseen test data and prove particularly useful for the image annotation task with multi-label scenarios in efficiently fusing visual and textual features.

In our future work, we envisage further investigation of the interplay between textual and visual content, in studying in particular the visual relatedness in regard to textual concepts. We also want to study some mid-level visual features or representations, for instance using an attentional model, which better account for affect related concepts.

Acknowledgement

This work was supported in part by the French research agency ANR through the VideoSense project under the grant 2009 CORD 026 02.

References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1349–1380.

[2] A. Mojsilović, J. Gomes, B. Rogowitz, Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues, Int. J. Comput. Vision 56 (2004) 79–107.

[3] J. Li, J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Trans. Pattern Anal. Mach. Intell. (2003) 1075–1088.

[4] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, TOMCCAP (2006) 1–19.

[5] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vision (2010) 303–338.

[6] A. F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and trecvid, in: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 321–330.

[7] M. J. Huiskes, M. S. Lew, M. S. Lew, The mir flickr retrieval evaluation, in: Multimedia Information Retrieval, 2008, pp. 39–43.

[8] M. J. Huiskes, B. Thomee, M. S. Lew, New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative, in: MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval, 2010, pp. 527–536.

[9] S. Nowak, K. Nagel, J. Liebetrau, The clef 2011 photo annotation and concept-based retrieval tasks, in: CLEF Workshop Notebook Paper, 2011.

[10] S. Nowak, M. J. Huiskes, New strategies for image annotation: Overview of the photo annotation task at imageclef 2010, in: CLEF Workshop Notebook Paper, 2010.

[11] G. Wang, D. Hoiem, D. A. Forsyth, Building text features for object image classification., in: Computer Vision Pattern Recognition, 2009, pp. 1367–1374.

[12] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: International Conference Computer Vision, 2003, pp. 1470–1477.

[13] M. Guillaumin, J. J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification., in: Computer Vision Pattern Recognition, 2010, pp. 902–909.

[14] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision (2004) 91–110.

[15] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1582–1596.

[16] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision Pattern Recognition, Vol. 2, 2005, pp. 886–893.

[17] C. Zhu, C.-E. Bichot, L. Chen, Visual object recognition using daisy descriptor, in: International Conference Multimedia and Expo., 2011, pp. 1–6.

[18] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition (1996) 51–59.

[19] C. Zhu, C.-E. Bichot, L. Chen, Multi-scale color local binary patterns for visual object classes recognition, in: ICPR, 2010, pp. 3065–3068.

[20] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision, European Computer Conference Vision, 2004, pp. 1–22.

[21] P. Valdez, A. Mehrabian, Effects of color on emotions, J. Exp Psychol Gen 123 (1994) 394–409.

[22] C. Colombo, A. D. Bimbo, P. Pala, Semantics in visual information retrieval, IEEE Multimedia 6 (1999) 38–53.

[23] H. Tamura, S. Mori, T. Yamawaki, Texture features corresponding to visual perception, IEEE Trans. on System, Man and Cybernatic 6 (1978) 460–473.

[24] N. Liu, E. Dellandréa, B. Tellez, L. Chen, L. Chen, Associating textual features with visual ones to improve affective image classification, in: Affective Computing Intelligent Interactive, 2011, pp. 195–204.

[25] N. Liu, E. Dellandréa, B. Tellez, L. Chen, Evaluation of Features and Combination Approaches for the Classification of Emotional Semantics in Images, in: International Conference on Computer Vision, Theory and Applications, 2011.

[26] R. Datta, J. Li, J. Z. Wang, Content-based image retrieval: approaches and trends of the new age, in: Multimedia Information Retrieval, 2005, pp. 253–262.

[27] Y. Ke, X. Tang, F. Jing, The design of high-level features for photo quality assessment, in: Computer Vision Pattern Recognition, Vol. 1, 2006, pp. 419–426.

[28] G. Salton, A. Wong, C. S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (1975) 613–620.

[29] S. T. Dumais, Latent semantic analysis, Tech. rep. (2005).

[30] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50–57.

[31] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[32] J. Fishbein, C. Eliasmith, Integrating structure and meaning: A new method for encoding structure for text classification, in: Advances in Information Retrieval, Vol. 4956 of Lecture Notes in Computer Science, 2008, pp. 514–521.

[33] A. Lavelli, F. Sebastiani, R. Zanoli, Distributional term representations: an experimental comparison, in: Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004, pp. 615–624.

[34] A. Moschitti, R. Basili, Complex linguistic features for text classification: a comprehensive study, in: Proceedings of the 26th European Conference on Information Retrieval, 2004, pp. 181–196.

[35] M. Sahlgren, R. Cöster, Using bag-of-concepts to improve the performance of support vector machines in text categorization, in: Proceedings of the 20th international conference on Computational Linguistics, 2004.

[36] H. J. Escalante, M. Montes, E. Sucar, Multimodal indexing based on semantic cohesion for image retrieval, Information Retrieval 15 (2011) 1–32.

[37] T. Mensink, G. Csurka, F. Perronnin, J. Snchez, J. J. Verbeek, Lear and xrce's participation to visual concept detection task - imageclef 2010, in: CLEF Workshop Notebook Paper, 2010.

[38] A. Binder, W. Samek, M. Kloft, C. Müller, K.-R. Müller, M. Kawanabe, The joint submission of the tu berlin and fraunhofer first (tubfi) to the imageclef2011 photo annotation task, in: CLEF Workshop Notebook Paper, 2011.

[39] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th annual ACM international conference on Multimedia, 2005, pp. 399–402.

[40] V. Parshin, A. Paradzinets, L. Chen, Multimodal data fusion for video scene segmentation, in: Visual Information and Information Systems, Vol. 3736 of Lecture Notes in Computer Science, 2006, pp. 279–289.

[41] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, J.-M. Renders, Crossing textual and visual content in different application scenarios, Multimedia Tools and Applications 42 (2009) 31–56.

[42] M. Worring, C. G. M. Snoek, B. Huurnink, J. C. van Gemert, D. C. Koelma, O. de Rooij, The mediamill large.lexicon concept suggestion engine, in: Proceedings of the 14th annual ACM international conference on Multimedia, 2006, pp. 785–786.

[43] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Koelma, F. J. Seinstra, The mediamill trecvid 2004 semantic video search engine, in: Proceedings of the TRECVID Workshop, 2004.

[44] T. Westerveld, A. P. D. Vries, A. van Ballegooij, F. de Jong, D. Hiemstra, A probabilistic multimedia retrieval model and its evaluation, EURASIP Journal on Applied Signal Processing 2003 (2003) 186–198.

[45] Y. Wu, E. Y. Chang, K. C.-C. Chang, J. R. Smith, Optimal multimodal fusion for multimedia data analysis, in: Proceedings of the 12th annual ACM international conference on Multimedia, 2004, pp. 572–579.

[46] A. Znaidia, H. L. Borgne, A. Popescu, Cea list's participation to visual concept detection task of imageclef 2011, in: CLEF Workshop Notebook Paper, 2011.

[47] B. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, J. Smith, Normalized classifier fusion for semantic visual concept detection, in: Int. ICIP, Vol. 2, 2003, pp. 535–538.

[48] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226–239.

[49] M. M. Bradley, P. J. Lang, Affective norms for english words (anew): instruction manual and affective ratings, Tech. rep., Center for Research in Psychophysiology, University of Florida (1999).

[50] D. Schwab, M. Lafourcade, V. Prince, Antonymy and conceptual vectors, in: Proceedings of the 19th international conference on Computational linguistics, Vol. 1, 2002, pp. 1–7.

[51] K. Scherer, Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, USA, 2001.

[52] M. Bradley, P. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, Journal of behavior therapy and experimental psychiatry 25 (1994) 49–59.

[53] A. Budanitsky, G. Hirst, Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures, in: Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, 2001.

[54] G. A. Miller, Wordnet: A lexical database for english, Communications of the ACM 38 (1995) 39–41.

[55] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, Int. J. Comput. Vision (2004) 63–86.

[56] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision Pattern Recognition, Vol. 2, 2006, pp. 2169–2178.

[57] F.-F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Computer Vision Pattern Recognition, 2005, pp. 524–531.

[58] E. Tola, V. Lepetit, P. Fua, Daisy: An efficient dense descriptor applied to wide-baseline stereo, IEEE Trans. Pattern Anal. Mach. Intell. (2010) 815–830.

[59] A. Pujol, L. Chen, Line segment based edge feature using hough transform, in: International Conference on Visualization, Imaging and Image Processing, 2007, pp. 201–206.

[60] P. Dunker, S. Nowak, A. Begau, C. Lanz, Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach, in: Multimedia Information Retrieval, 2008, pp. 97–104.

[61] J. Itten, The art of color: the subjective experience and objective rationale of color, New York: Reinhold Pub. Corp., 1961.

[62] J. Machajdik, A. Hanbury, Affective image classification using features inspired by psychology and art theory, in: ACM Multimedia, 2010, pp. 83–92.

[63] E. Dellandréa, N. Liu, L. Chen, Classification of affective semantics in images based on discrete and dimensional models of emotions, in: International Workshop on Content-Based Multimedia Indexing, 2010, pp. 99–104.

[64] R. O. Duda, P. E. Hart, Use of the hough transformation to detect lines and curves in pictures, Commun. ACM (1972) 11–15.

[65] P. A. Viola, M. J. Jones, Robust real-time face detection, in: International Conference Computer Vision, Vol. 57, 2001, pp. 137–154.

[66] R. M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE 67 (1979) 786–804.

[67] N. A. Anstey, Correlation techniques - a reivew, Canadian Journal of Exploration Geophysics 2 (1966) 55–82.

[68] K. van de Sande, University of Amsterdam, ColorDescriptor software, `http://www.colordescriptors.com`.

[69] W. Ben Soltana, D. Huang, M. Ardabilian, L. Chen, C. Ben Amar, Comparison of 2D/3D Features and Their Adaptive Score Level Fusion for 3D Face Recognition, in: 3D Data Processing, Visualization and Transmission, 2010.

[70] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (1994) 1119–1125.

[71] L. Rokach, Ensemble-based classifiers, Artificial Intelligence Review 33 (2010) 1–39.

[72] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.

[73] J. A. Russell, A circumplex model of affect, Journal of Personality and Social Psychology 39 (1980) 1161–1178.

[74] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer New York Inc., New York, NY, USA, 1995.

[75] J. Zhang, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, Int. J. Comput. Vision 73 (2007) 213–238.

[76] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1–27.